

Clustering and Variable Selection in the Presence of Mixed Variable Types and Missing Data

BY C. B. STORLIE[†], S. M. MYERS[‡], S. K. KATUSIC[†], A. L. WEAVER[†], R. VOIGT[§],
R. C. COLLIGAN[†], P. E. CROARKIN[†], R. E. STOECKEL[†], J. D. PORT[†]

[†]*Mayo Clinic* [‡]*Geisinger Autism & Developmental Medicine Institute*

[§]*Texas Children's Hospital*

SUMMARY

We consider the problem of model-based clustering in the presence of many correlated, mixed continuous and discrete variables, some of which may have missing values. Discrete variables are treated with a latent continuous variable approach and the Dirichlet process is used to construct a mixture model with an unknown number of components. Variable selection is also performed to identify the variables that are most influential for determining cluster membership. The work is motivated by the need to cluster patients thought to potentially have autism spectrum disorder (ASD) on the basis of many cognitive and/or behavioral test scores. There are a modest number of patients (~ 480) in the data set along with many (~ 100) test score variables (many of which are discrete valued and/or missing). The goal of the work is to (i) cluster these patients into similar groups to help identify those with similar clinical presentation, and (ii) identify a sparse subset of tests that inform the clusters in order to eliminate unnecessary testing. The proposed approach compares very favorably to other methods via simulation of problems of this type. The results of the ASD analysis suggested three clusters to be most likely, while only four test scores had high (> 0.5) posterior probability of being informative. This will result in much more efficient and informative testing. The need to cluster observations on the basis of many correlated, continuous/discrete variables with missing values, is a common problem in the health sciences as well as in many other disciplines.

Some key words: Model-Based Clustering; Dirichlet Process; Missing Data; Hierarchical Bayesian Modeling; Mixed Variable Types; Variable Selection.

1. INTRODUCTION

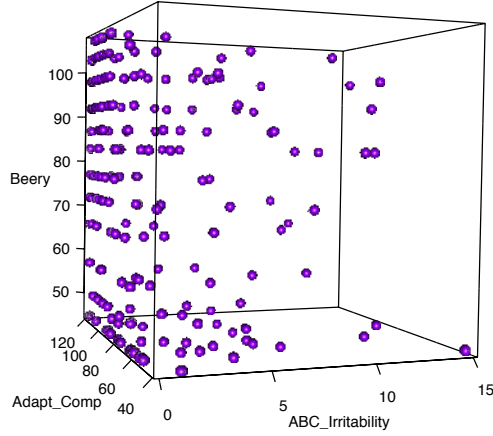
Model-based clustering has become a very popular means for unsupervised learning (Fraley and Raftery, 2002; Basu and Chib, 2003; Quintana and Iglesias, 2003; Tadesse et al., 2005). This is due in part to the ability to use the model likelihood to inform, not only the cluster membership, but also the number of clusters M which has been a heavily researched problem for many years. The most widely used model-based approach is the normal mixture model which is not suitable for mixed continuous/discrete variables. For example, this work is motivated by the need to cluster patients thought to potentially have autism spectrum disorder (ASD) on the basis of many correlated test scores. There are a modest number of patients (~ 480) in the training data set along with many (~ 100) test score/self-report variables, many of which are discrete valued or have left or right boundaries. Figure 1 provides a look at the data across three of the variables; Beery_standard is discrete valued and ABC_irritability is continuous, but with significant mass at the left boundary of zero. The goals of this problem are to (i) cluster these patients into similar groups to help identify those with similar clinical presentation, and (ii) identify a sparse subset of tests that inform the clusters in an effort to eliminate redundant testing. This problem is also complicated by the fact that many patients in the training data have missing test scores. The need to cluster incomplete observations on the basis of many correlated continuous/discrete variables is a common problem in the health sciences as well as in many other disciplines.

When clustering in high dimensions, it becomes critically important to use some form of dimension reduction or variable selection to achieve accurate cluster formation. A common approach to deal with this is a principle components or factor approach (e.g., Liu et al., 2003). However, such a solution does not address goal (ii) above for the ASD clustering problem. The problem of variable selection in regression or conditional density estimation has been well studied from both the L_1 penalization (see, e.g., Tibshirani, 1996; Zou and Hastie, 2005; Lin and Zhang, 2006) and Bayesian perspectives (George and McCulloch, 1993; Reich et al., 2009; Chung and Dunson, 2012). However, variable selection in clustering is more challenging than that in regression as there is no response to guide (supervise) the selection. Still, there have been several articles considering this topic. For example, Raftery and Dean (2006) propose a partition of the variables into *informative* (dependent on cluster membership even after conditioning on all of the other variables) and *non-informative* (conditionally independent of cluster membership given the values of the other variables). They use BIC to accomplish variable selection with a greedy search. Similar approaches are used by Maugis et al. (2009) and Fop et al. (2015). The popular LASSO or L_1 type penalization has also been applied to shrink cluster means together for variable selection (Pan and Shen, 2007; Wang and Zhu, 2008; Xie et al., 2008). There have also been several approaches developed for sparse K-means and distance based clustering (Friedman and Meulman, 2004; Hoff, 2006; Witten and Tibshirani, 2012).

In the Bayesian literature Tadesse et al. (2005) consider variable selection in the finite normal mixture model using reversible jump (RJ) Markov chain Monte Carlo (MCMC) (Richardson and Green, 1997). Kim et al. (2006) extend that work to the nonparametric Bayesian mixture model via the Dirichlet process model (DPM) (Ferguson, 1973; Neal, 2000; Teh et al., 2006; Lid Hjort et al., 2010). The DPM has the advantage of allowing for a countably infinite number of possible components, while providing a posterior distribution for how many components have been *observed* in the data set at hand. Both Tadesse et al. (2005) and Kim et al. (2006) use a point mass prior to achieve sparse representation of the informative variables. However, for convenience they assume all non-informative variables are (unconditionally) independent of the informative variables. This assumption is frequently violated in practice and it is particularly problematic in the case of the ASD analysis as it would force far too many variables to be included into the informative set as is demonstrated later in this paper.

There is not a generally accepted best practice to clustering with mixed discrete and continuous variables. Hunt and Jorgensen (2003) and Murray and Reiter (2016) meld mixtures of *independent* multinomials for the categorical variables and mixtures of Gaussian for the continuous variables. However, it may not be desirable for the dependency between the discrete variables to be entirely represented by mixture components when clustering is the primary objective. As pointed out in Hennig and Liao (2013), mixture models can approximate any distribution arbitrarily well so care must be taken to ensure the mixtures fall in line with the goals of clustering. When using mixtures of Gaussian combined with independent multinomials, a data set with many correlated discrete variables will tend to result in more clusters than a comparable dataset with mostly continuous variables. A discrete variable measure of some quantity instead of the continuous version could therefore result in very different clusters. Thus, a Gaussian latent vari-

Fig. 1: 3D scatter plot of three of the test score variables for potential ASD subjects.



able approach (Muthen, 1983; Dunson, 2000) would seem more appropriate for treating discrete variables when clustering is the goal. An observed ordinal variable x_j , for example, is assumed to be the result of thresholding a latent Gaussian variable z_j . For binary variables, this reduces to the multivariate probit model (Lesaffre and Molenberghs, 1991; Chib and Greenberg, 1998). There are also extensions of this approach to allow for unordered categorical variables.

In this paper, we propose a Bayesian nonparametric approach to perform simultaneous estimation of the number of clusters, cluster membership, and variable selection while explicitly accounting for discrete variables and partially observed data. To the best of our knowledge, this is the first model-based clustering approach to allow for this complex but common data structure. The discrete variables as well as continuous variables with boundaries are treated with a Gaussian latent variable approach. The informative variable construct of Raftery and Dean (2006) for normal mixtures is then adopted. However, in order to effectively handle the missing values and account for uncertainty in the variable selection and number of clusters, the proposed model is cast in a fully Bayesian framework via the Dirichlet process. This is then similar to the work of Kim et al. (2006), however, they did not consider discrete variables or missing data. Further, a key result of this paper is a solution to allow for dependence between informative and non-informative variables in the nonparametric Bayesian mixture model. This solution takes a particularly simple form and also provides an intuitive means with which to define the prior distribution in a manner that decreases prior sensitivity. The component parameters are marginalized out to facilitate more efficient MCMC sampling via a modified version of the split-merge algorithm of Jain and Neal (2004). Finally, missing data is then handled in a principled manner by treating missing values as unknown parameters in the Bayesian framework (see, e.g., Storlie et al., 2015, 2017). This approach implicitly assumes a missing at random (MAR) mechanism (Rubin, 1976), which implies that the likelihood of a missing value *can* depend on the value of the unobserved variable(s), marginally, just not after conditioning on the observed variables.

The rest of the paper is laid out as follows. Section 2 describes the proposed nonparametric Bayesian approach to clustering with mixed discrete and continuous variables with variable selection. Section 3 evaluates the performance of this approach when compared to other methods on several simulation cases. The approach is then applied to the problem for which it was designed in Section 4 where a comprehensive analysis of the ASD problem is presented. Section 5 concludes the paper. This paper also has supplementary material which contains likelihood derivations, full exposition of the MCMC algorithm used to fit the model, and MCMC trace plots.

2. METHODOLOGY

2.1. Dirichlet Process Mixture Models

As discussed above, the proposed model for clustering uses mixture distributions with a countably infinite number of components via the Dirichlet process prior (Ferguson, 1973; Escobar and West, 1995; MacEachern and Müller, 1998). Let $\mathbf{y} = (y_1, \dots, y_p)$ be a p -variate random vector and let \mathbf{y}_i , $i = 1, \dots, n$, denote the i^{th} observation of \mathbf{y} . It is assumed that \mathbf{y}_i are independent random vectors coming from distribution $F(\theta_i)$. The model parameters θ_i are assumed to come from a mixing distribution G which has a Dirichlet process prior, i.e., the familiar model,

$$\mathbf{y}_i \mid \theta_i \sim F(\theta_i), \quad \theta_i \sim G, \quad G \sim \text{DP}(G_0, \alpha), \quad (1)$$

where DP represents a Dirichlet Process distribution, G_0 is the base distribution and α is a precision parameter, determining the concentration of the prior for G about G_0 (Escobar and West, 1995). The prior distribution for θ_i in terms of successive conditional distributions is obtained by integrating over G , i.e.,

$$\theta_i \mid \theta_1, \dots, \theta_{i-1} \sim \frac{1}{i-1+\alpha} \sum_{i'=1}^{i-1} \delta(\theta_{i'}) + \frac{\alpha}{i-1+\alpha} G_0, \quad (2)$$

where $\delta(\theta)$ is a point mass distribution at θ . The representation in (2) makes it clear that (1) can be viewed as a countably infinite mixture model. Alternatively, let $\Omega = [\omega_1, \omega_2, \dots]$ denote the unique values of the θ_i and let ϕ_i be the index for the component to which observation i belongs, i.e., so that $\omega_{\phi_i} = \theta_i$. The following model (Neal, 2000) is equivalent to (2)

$$P(\phi_i = m \mid \phi_1, \dots, \phi_{i-1}) = \begin{cases} 1 & \text{if } i = 1 \text{ and } m = 1. \\ \frac{n_{i,m}}{i-1+\alpha} & \text{if } \phi_{i'} = m \text{ for any } i' < i. \\ \frac{\alpha}{i-1+\alpha} & \text{if } m = \max(\phi_1, \dots, \phi_{i-1}) + 1. \\ 0 & \text{otherwise,} \end{cases} \quad (3)$$

with $\mathbf{y}_i \mid \phi_i, \Omega \sim F(\omega_{\phi_i})$, $\omega_m \sim G_0$ and $n_{i,m}$ is the number of $\phi_{i'} = m$ for $i' < i$. Thus, a new observation i is allocated to an existing cluster with probability proportional to the cluster size or it is assigned to a new cluster with probability proportional to α . This is often called the Chinese restaurant representation of the Dirichlet process. It is often assumed that F is a Gaussian distribution in which case $\omega_m = (\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ describes the mean and covariance for the m^{th} component. This results in a normal mixture model with a countably infinite number of components.

2.2. Discrete Variables and Boundaries/Censoring

Normal mixture models are not effective for clustering when some of the variables are too discretized as demonstrated in Section 3. This is also a problem when the data have left or right boundaries that can be achieved (e.g., several people score the minimum or maximum on a test). However, a Gaussian latent variable approach can be used to circumvent these issues. Suppose that variables y_j for $j \in \mathcal{D}$ are discrete, ordinal variables taking on possible values $\mathbf{d}_j = \{d_{j,1}, \dots, d_{j,L_j}\}$ and that y_j for $j \in \mathcal{C} = \mathcal{D}^c$ are continuous variables with lower and upper limits of b_j and c_j , which could be infinite. Assume for some latent, p -variate, continuous random vector \mathbf{z} that

$$y_j = \begin{cases} \sum_{l=1}^{L_j} d_{j,l} I_{\{a_{j,l-1} < z_j \leq a_{j,l}\}} & \text{for } j \in \mathcal{D} \\ z_j I_{\{b_j \leq z_j \leq c_j\}} + b_j I_{\{z_j < b_j\}} + c_j I_{\{z_j > c_j\}} & \text{for } j \in \mathcal{C} \end{cases} \quad (4)$$

where I_A is the indicator function equal to 1 if A and 0 otherwise, $a_{j,0} = -\infty$, $a_{j,L_j} = \infty$, and $a_{j,l} = d_{j,l}$ for $l = 1, \dots, L_j - 1$. That is, the discrete y_j are the result of thresholding the latent variable z_j on the respective cut-points. The continuous y_j variables are simply equal to the z_j unless the z_j cross the left or right boundary of what can be observed for y_j . That is, if there are finite limits for y_j , then y_j is assumed to be a left and/or right censored version of z_j , thus producing a positive mass at the boundary values of y_j .

A joint mixture model for mixed discrete and continuous variables can then be represented as

$$\mathbf{z}_i \mid \phi_i, \Omega \sim N(\boldsymbol{\mu}_{\phi_i}, \boldsymbol{\Sigma}_{\phi_i}), \quad (5)$$

with prior distributions for ω_m and $\boldsymbol{\phi} = [\phi_1, \dots, \phi_n]'$ as in (3).

Binary y_j such as gender can be accommodated by setting $\mathbf{d}_j = \{0, 1\}$. However, if there is only one cut-point then the model must be restricted for identifiability (Chib and Greenberg, 1998); namely, if y_j is binary, then we must set $\boldsymbol{\Sigma}_m(j, j) = 1$. The restriction that $\boldsymbol{\Sigma}_m(j, j) = 1$ for binary y_j complicates posterior inference, however, this problem has been relatively well studied in the multinomial probit and multivariate probit settings and various proposed solutions exist (McCulloch et al., 2000; Imai and van Dyk, 2005). It is also straight-forward to use the latent Gaussian variable approach to allow for unordered categorical variables (McCulloch et al., 2000; Imai and van Dyk, 2005; Zhang et al., 2008; Bhattacharya and Dunson, 2012), however, inclusion of categorical variables also complicates notation and there are no such variables in the ASD data. For brevity, attention is restricted here to continuous and ordinal discrete variables.

2.3. Variable Selection

Variable selection in clustering problems is more challenging than in regression problems due to the lack of targeted information with which to guide the selection. Using a model-based clustering approach allows a likelihood based approach to model selection, but exactly how the parameter space should be restricted when a variable is “out of the model” requires some care. Raftery and Dean (2006) defined a variable y_j to be *non-informative* if conditional on the values of the other variables, it is independent of cluster membership. This implies that a non-informative y_j may still be quite dependent on cluster membership through its dependency with other variables. They assumed a Gaussian mixture distribution for the informative variables, with a conditional Gaussian distribution for the non-informative variables and used maximum likelihood to obtain the change in BIC between candidate models. Thus, they accomplished variable selection with a greedy search to minimize BIC. They further considered restricted covariance parameterizations to reduce the parameter dimensionality (e.g., diagonal, common volume, common shape, common orientation, and combinations of these restrictions). We instead take a Bayesian approach to this problem via Stochastic Search Variable Selection (SSVS) (George and McCulloch, 1993, 1997) as this allows for straight-forward treatment of uncertainty in the selected variables and that due to missing values. Kim et al. (2006) used such an approach with a DPM for infinite normal mixtures, however, due to the difficulty imposed they did not use the same definition as Raftery and Dean (2006) for a non-informative variable. They defined a non-informative variable to be one that is (unconditionally) independent of cluster membership and all other variables. This is not reasonable in many cases, particularly in the ASD problem, and can result in negative consequences as seen in Section 3. Below, we layout a more flexible model specification akin to that taken in Raftery and Dean (2006) to allow for dependence between informative and non-informative variables in the nonparametric Bayesian mixture model.

Let the informative variables be represented by the *model* γ , a vector of binary values such that $\{y_j : \gamma_j = 1\}$ is the set of informative variables. A priori it is assumed that $\text{pr}(\gamma_j = 1) = \rho_j$. Without loss of generality assume that \mathbf{y} has elements ordered such that $\mathbf{y} = [\mathbf{y}^{(1)}, \mathbf{y}^{(2)}]$, with $\mathbf{y}^{(1)} = \{y_j : \gamma_j = 1\}$ and $\mathbf{y}^{(2)} = \{y_j : \gamma_j = 0\}$, and similarly for $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$. The model in (5) becomes,

$$\mathbf{z}_i \mid \gamma, \phi_i, \Omega \sim N(\boldsymbol{\mu}_{\phi_i}, \boldsymbol{\Sigma}_{\phi_i}), \quad (6)$$

with

$$\boldsymbol{\mu}_m = \begin{pmatrix} \boldsymbol{\mu}_{m1} \\ \boldsymbol{\mu}_{m2} \end{pmatrix}, \quad \boldsymbol{\Sigma}_m = \begin{pmatrix} \boldsymbol{\Sigma}_{m11} & \boldsymbol{\Sigma}_{m12} \\ \boldsymbol{\Sigma}_{m21} & \boldsymbol{\Sigma}_{m22} \end{pmatrix}. \quad (7)$$

Then from standard multivariate normal theory, $\mathbf{z}^{(2)} \mid \mathbf{z}^{(1)}, \phi = m \sim N(\boldsymbol{\mu}_{2|1}, \boldsymbol{\Sigma}_{2|1})$ with $\boldsymbol{\mu}_{2|1} = \boldsymbol{\mu}_{m2} + \boldsymbol{\Sigma}_{m21} \boldsymbol{\Sigma}_{m11}^{-1} (\mathbf{z}^{(1)} - \boldsymbol{\mu}_{m1})$ and $\boldsymbol{\Sigma}_{22|1} = \boldsymbol{\Sigma}_{m22} - \boldsymbol{\Sigma}_{m21} \boldsymbol{\Sigma}_{m11}^{-1} \boldsymbol{\Sigma}_{m12}$. Now in order for the non-informative variables to follow the definition of Raftery and Dean (2006), the $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ must be parameterized so that $\boldsymbol{\mu}_{2|1}, \boldsymbol{\Sigma}_{2|1}$ do not depend on m . In order to accomplish this, it is helpful to make use of the canonical parameterization of the Gaussian (Rue and Held, 2005),

$$\mathbf{z} \mid \gamma, \Omega, \phi = m \sim \mathcal{N}_C(\mathbf{b}_m, \mathbf{Q}_m),$$

with precision $\mathbf{Q}_m = \boldsymbol{\Sigma}_m^{-1}$ and $\mathbf{b}_m = \mathbf{Q}_m \boldsymbol{\mu}_m$. Partition the canonical parameters as,

$$\mathbf{b}_m = \begin{pmatrix} \mathbf{b}_{m1} \\ \mathbf{b}_2 \end{pmatrix}, \quad \mathbf{Q}_m = \begin{pmatrix} \mathbf{Q}_{m11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{pmatrix}. \quad (8)$$

Result 1. The parameterization in (8) results in $(\boldsymbol{\mu}_{2|1}, \boldsymbol{\Sigma}_{2|1})$ that does not depend on m .

Proof. The inverse of a partitioned matrix directly implies that $\boldsymbol{\Sigma}_{2|1} = \mathbf{Q}_{22}^{-1}$, which does not depend on m . It also implies that $-\mathbf{Q}_{22}^{-1} \mathbf{Q}_{21} = \boldsymbol{\Sigma}_{m21} \boldsymbol{\Sigma}_{m11}^{-1}$, and substituting $\boldsymbol{\Sigma}_m \mathbf{b}_m$ for $\boldsymbol{\mu}_m$ in

$\mu_{2|1}$ gives $\mu_{2|1} = \mathbf{Q}_{22}^{-1} (\mathbf{b}_2 - \mathbf{Q}_{21} \mathbf{z}^{(1)})$, which also does not depend on m . \square

Now the problem reduces to defining a prior distribution for Ω , i.e., $\omega_m = \{\mathbf{b}_m, \mathbf{Q}_m\}$, $m = 1, 2, \dots$, conditional on the model γ , that maintains the form of (8). Let $\omega_m^{(1)} = \{\mathbf{b}_{m1}, \mathbf{Q}_{m11}\}$ and $\omega_m^{(2)} = \omega^{(2)} = \{\mathbf{b}_2, \mathbf{Q}_{21}, \mathbf{Q}_{22}\}$. The prior distribution for Ω will be defined first unconditionally for $\omega^{(2)}$ and then for $\omega_m^{(1)}$, $m = 1, 2, \dots$, conditional on $\omega^{(2)}$. There are several considerations in defining these distributions: (i) the resulting \mathbf{Q}_m must be positive definite, (ii) it is desirable for the marginal distribution of (μ_m, Σ_m) to remain unchanged for any model γ to limit the influence of the prior for ω_m on variable selection, and (iii) it is desirable for them to be conjugate to facilitate MCMC sampling (Neal, 2000; Jain and Neal, 2004). Let Ψ be a $p \times p$ positive definite matrix, partitioned just as \mathbf{Q}_m , and for a given γ assume the following distribution for $\omega^{(2)}$,

$$\begin{aligned} \mathbf{Q}_{22} &\sim \mathcal{W}(\Psi_{22|1}^{-1}, \eta), \quad \mathbf{b}_2 | \mathbf{Q}_{22} \sim \mathcal{N}(\mathbf{0}, \frac{1}{\lambda} \mathbf{Q}_{22}), \\ \mathbf{Q}_{21} | \mathbf{Q}_{22} &\sim \mathcal{MN}(-\mathbf{Q}_{22} \Psi_{21} \Psi_{11}^{-1}, \mathbf{Q}_{22}, \Psi_{11}^{-1}), \end{aligned} \quad (9)$$

where \mathcal{W} denotes the Wishart distribution, and \mathcal{MN} denotes the matrix normal distribution. The distribution of $\omega_m^{(1)}$, conditional on $\omega^{(2)}$ is then defined implicitly below in terms of (μ_{m1}, Σ_{m11}) ,

$$\Sigma_{m11} \stackrel{iid}{\sim} \mathcal{W}^{-1}(\Psi_{11}, \eta - p_2), \quad \mu_{m1} | \Sigma_{m11} \stackrel{ind}{\sim} \mathcal{N}(\mathbf{0}, \frac{1}{\lambda} \Sigma_{m11}), \quad (10)$$

where \mathcal{W}^{-1} denotes the inverse-Wishart distribution and (μ_{m1}, Σ_{m11}) are independent of $\omega^{(2)}$. This is *not* to say that $\omega_m^{(1)} = (\mathbf{b}_{m1}, \mathbf{Q}_{m11})$ is independent of $\omega^{(2)}$, rather the distribution imposed on $(\mathbf{b}_{m1}, \mathbf{Q}_{m11})$ via (9) and (10) is quite dependent on $\omega^{(2)}$ via the relations, $\mathbf{b}_{m1} = \Sigma_{m11}^{-1} \mu_{m1} + \mathbf{Q}_{12} \mathbf{Q}_{22}^{-1} \mathbf{b}_2$, and $\mathbf{Q}_{m11} = \Sigma_{m11}^{-1} + \mathbf{Q}_{12} \mathbf{Q}_{22}^{-1} \mathbf{Q}_{21}$.

Result 2. The prior distribution defined in (9) and (10) results in a marginal distribution for (μ_m, Σ_m) of $\mathcal{NIW}(\mathbf{0}, \lambda, \Psi, \eta)$, i.e., the same normal-inverse-Wishart regardless of γ .

Proof. It follows from Theorem 3 of Bodnar and Okhrin (2008) that $\Sigma_m \sim \mathcal{IW}(\eta, \Psi)$. It remains to show $\mu_m | \Sigma_m \sim \mathcal{N}(\mathbf{0}, (1/\lambda) \Sigma_m)$. However, according to (9) and (10) and the independence assumption,

$$\begin{pmatrix} \mu_{m1} \\ \mathbf{b}_2 \end{pmatrix} \Big| \Sigma_m \sim \mathcal{N} \left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \frac{1}{\lambda} \begin{pmatrix} \Sigma_{m11} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_{22} \end{pmatrix} \right).$$

Also, $\mathbf{b}_m = \mathbf{Q}_m \mu_m$ implies,

$$\begin{pmatrix} \mu_{m1} \\ \mu_{m2} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{Q}_{22}^{-1} \mathbf{Q}_{21} & \mathbf{Q}_{22}^{-1} \end{pmatrix} \begin{pmatrix} \mu_{m1} \\ \mathbf{b}_2 \end{pmatrix}.$$

Using the relation $\mathbf{A}x \sim \mathcal{N}(\mathbf{A}\mu, \mathbf{A}\Sigma\mathbf{A}')$ for $x \sim \mathcal{N}(\mu, \Sigma)$ gives the desired result. \square

As mentioned above, the normal-inverse-Wishart distribution is conjugate for ω_m in the unrestricted (no variable selection) setting. It turns out that the distribution defined in (9) and (10) is conjugate for the parameterization in (8) as well, so that the component parameters can be integrated out of the likelihood. Let the (latent) observations be denoted as $\mathbf{Z} = [\mathbf{z}'_1, \dots, \mathbf{z}'_n]'$, and the data likelihood as $f(\mathbf{Z} | \gamma, \phi, \Omega)$.

Result 3. The marginal likelihood of \mathbf{Z} is given by

$$\begin{aligned} f(\mathbf{Z} | \gamma, \phi) &= \int f(\mathbf{Z} | \gamma, \phi, \Omega) f(\Omega | \gamma) d\Omega \\ &= \pi^{-\frac{np}{2}} \prod_{m=1}^M \left[\left(\frac{\lambda}{n_m + \lambda} \right)^{\frac{p_1}{2}} \frac{|\Psi_{11}|^{\frac{\eta - p_2}{2}} \Gamma_{p_1}(\frac{n_m + \eta - p_2}{2})}{|\mathbf{V}_{m11}|^{\frac{\eta - p_2}{2}} \Gamma_{p_1}(\frac{\eta - p_2}{2})} \right] \left[\left(\frac{\lambda}{n + \lambda} \right)^{\frac{p_2}{2}} \frac{|\Psi_{11}|^{\frac{p_2}{2}} |\Psi_{21}|^{\frac{\eta}{2}} \Gamma_{p_2}(\frac{n + \eta}{2})}{|\mathbf{V}_{11}|^{\frac{p_2}{2}} |\mathbf{V}_{21}|^{\frac{n + \eta}{2}} \Gamma_{p_2}(\frac{\eta}{2})} \right], \end{aligned}$$

where (i) $M = \max(\phi)$, i.e., the number of observed components, (ii) $p_1 = \sum \gamma_j$ is the number of informative variables, (iii) $p_2 = p - p_1$, (iv) n_m is the number of $\phi_i = m$, (v) $\Gamma_p(\cdot)$ is the multivariate gamma function, and (vi) \mathbf{V}_{m11} , \mathbf{V}_{11} , $\mathbf{V}_{2|1}$ are defined as,

$$\begin{aligned} \mathbf{V}_{m11} &= \sum_{\phi_i=m} (\mathbf{z}_i^{(1)} - \bar{\mathbf{z}}_{m1})(\mathbf{z}_i^{(1)} - \bar{\mathbf{z}}_{m1})' + \frac{n_m \lambda}{n_m + \lambda} \bar{\mathbf{z}}_{m1} \bar{\mathbf{z}}_{m1}' + \boldsymbol{\Psi}_{11}, \\ \mathbf{V}_{11} &= \sum_{i=1}^n (\mathbf{z}_i^{(1)} - \bar{\mathbf{z}}_1)(\mathbf{z}_i^{(1)} - \bar{\mathbf{z}}_1)' + \frac{n \lambda}{n + \lambda} \bar{\mathbf{z}}_1 \bar{\mathbf{z}}_1' + \boldsymbol{\Psi}_{11}, \quad \mathbf{V}_{2|1} = \mathbf{V}_{22} - \mathbf{V}_{21} \mathbf{V}_{11}^{-1} \mathbf{V}_{21}', \\ \text{with } \bar{\mathbf{z}}_{m1} &= \frac{1}{n_m} \sum_{\phi_i=m} \mathbf{z}_i^{(1)}, \quad \bar{\mathbf{z}}_1 = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i^{(1)}, \quad \bar{\mathbf{z}}_2 = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i^{(2)}, \\ \mathbf{V}_{22} &= \sum_{i=1}^n (\mathbf{z}_i^{(2)} - \bar{\mathbf{z}}_2)(\mathbf{z}_i^{(2)} - \bar{\mathbf{z}}_2)' + \frac{n \lambda}{n + \lambda} \bar{\mathbf{z}}_2 \bar{\mathbf{z}}_2' + \boldsymbol{\Psi}_{22}, \text{ and} \\ \mathbf{V}_{21} &= \sum_{i=1}^n (\mathbf{z}_i^{(2)} - \bar{\mathbf{z}}_2)(\mathbf{z}_i^{(1)} - \bar{\mathbf{z}}_1)' + \frac{n \lambda}{n + \lambda} \bar{\mathbf{z}}_2 \bar{\mathbf{z}}_1' + \boldsymbol{\Psi}_{21}. \end{aligned} \tag{11}$$

The derivation of Result 3 is provided in the Supplementary Material.

2.4. Hyper-Prior Distributions

Kim et al. (2006) found there to be a lot of prior sensitivity due to the choice of prior for the component parameters. This is in part due to the separate prior specification for the parameters corresponding to informative and non-informative variables, respectively. The specification above treats all component parameters collectively, in a single prior, so that the choice will not be sensitive to the interplay between the priors chosen for informative and non-informative variables. A further stabilization can be obtained by rationale similar to that used in Raftery and Dean (2006) for restricted forms of the covariance (such as equal shape, orientation, etc.). We do not enforce such restrictions exactly, but one might expect the components to have similar covariances or similar means for some of the components. Thus it makes sense to put hierarchical priors on λ , $\boldsymbol{\Psi}$, and η , to encourage such similarity if warranted by the data. A Gamma prior is also placed on the concentration parameter α , i.e.,

$$\begin{aligned} \lambda &\sim \text{Gamma}(A_\lambda, B_\lambda), \quad \eta - (p+1) \sim \text{Gamma}(A_\eta, B_\eta), \\ \boldsymbol{\Psi} &\sim \mathcal{W}(\mathbf{P}, N), \quad \alpha \sim \text{Gamma}(A_\alpha, B_\alpha). \end{aligned} \tag{12}$$

In the analyses below, relatively vague priors were used with $A_\lambda = B_\lambda = A_\eta = B_\eta = 2$. The prior for α was set to $A_\alpha = 2$, $B_\alpha = 10$, to encourage anywhere from 1 to 10 clusters from 100 observations. The results will still have some sensitivity to the choice of \mathbf{P} . In addition, there are well known issues with using a Wishart prior on variables of differing scale. In order to alleviate these issues, we recommend first standardizing the columns of the data to mean zero and unit variance, then using $N = p + 2$, $\mathbf{P} = (1/N)\mathbf{I}$. Finally, the prior probability for variable inclusion was set to $\rho_j = 0.5$ for all j . The data model defined in (4) and (6), the component prior distribution defined in (9) and (10), along with the hyper-priors in (12), completes the model specification.

2.5. MCMC Algorithm

Complete MCMC details are provided in the Supplementary Material. However, an overview is provided here to illustrate the main idea. The complete list of parameters to be sampled in the MCMC are $\Theta = \{\gamma, \phi, \lambda, \eta, \boldsymbol{\Psi}, \alpha, \tilde{\mathbf{Z}}\}$, where $\tilde{\mathbf{Z}}$ contains any latent element of \mathbf{Z} (i.e., either corresponding to missing data, discrete variable, or boundary/censored observation). Because the component parameters are integrated out, the ϕ_i can be updated with simple Gibbs sampling (Neal, 2000), however, this approach has known mixing issues (Jain and Neal, 2004; Ishwaran and James, 2011). Thus, a modified split-merge algorithm (Jain and Neal, 2004) similar to that

used in (Kim et al., 2006) was developed to sample from the posterior distribution of ϕ . The remaining parameters are updated in a hybrid Gibbs, Metropolis Hastings (MH) fashion. The γ vector is updated with MH by proposing an add, delete, or swap move (George and McCulloch, 1997). The \tilde{Z} are block updated, each with a MH step, but with a proposal that looks almost conjugate, and is therefore accepted with high probability; the block size can be adjusted to trade-off between acceptance and speed (e.g., acceptance $\sim 40\%$). A similar strategy is taken with the Ψ update, i.e., a nearly conjugate update is proposed and accepted/rejected via an MH step. The λ, η, α parameters have standard MH random walk updates on log-scale. The MCMC routine then consists of applying each of the above updates in turn to complete a single MCMC iteration, with the exception that the γ update be applied L_g times each iteration.

Two modifications were also made to the above computational strategy to improve mixing. The algorithm above would at times have trouble breaking away from local modes when proposing ϕ and γ updates separately. Thus, an additional joint update is proposed for ϕ and γ each iteration. Also, as described in more detail in the Supplementary Material, the traditional split merge algorithm proposes an update by first selecting two points, i and i' , at random. If they are from the same cluster (according to the current ϕ) it then assigns them to separate clusters and assigns the remaining points from that cluster to each of the two new clusters at random. It then conducts several (L) restricted (to one of the two clusters) Gibbs sampling updates to the remaining ϕ_h from the original cluster. The resulting ϕ^* then becomes the proposal for a split move. We found that the following adjustment resulted in much higher acceptance of split/merge moves. Instead of assigning the remaining points to the two clusters at random, simply assign them to the closest of the two observations i or i' . Then conduct L restricted Gibbs sample updates to produce the proposal. We found very little performance gain beyond $L = 3$.

It would also be possible to instead use a finite mixture approximation and sample via the kernel stick breaking representation of a DPM (Sethuraman, 1994; Ishwaran and James, 2011) to alleviate the slow mixing concerns with Gibbs updates for the ϕ_i . However, this approach would be complicated by the dependency between γ and the structure and dimensionality of the component parameters. This issue is entirely avoided with the proposed approach.

2.6. Inference for ϕ and γ

The estimated cluster membership $\hat{\phi}$ for all of the methods were taken to be the respective mode of the estimated cluster membership probability. For the DPM methods, the cluster membership probability matrix P (which is an $n \times \infty$ matrix in principle) is not sampled in the MCMC, and is not identified due to many symmetric modes (and thus their can be label switching in the posterior samples). However, the information theoretic approach of Stephens (2000) (applied to the DPM in Fu et al. (2013)) can be used to address this issue and relabel the posterior samples of ϕ to provide an estimate of P . The resulting estimate \hat{P} has i^{th} row, m^{th} column that can be thought of as the proportion of the relabeled posterior samples of ϕ_i that have the value m . While technically P is an $n \times \infty$ matrix, all columns after M^* have zero entries in \hat{P} , where M^* is the maximum number of clusters observed in the posterior. For the results below, the point estimate of the model $\hat{\gamma}$ is determined by $\hat{\gamma}_j = 1$ if $\text{pr}(\gamma_j = 1) > \rho_j = 0.5$.

3. SIMULATION RESULTS

In this section the performance of the proposed approach for clustering is evaluated on two simulation cases similar in nature to the ASD clustering problem. Each of the cases is examined (i) without missing data or discrete variables/censoring, (ii) with missing data, but no discrete variables/censoring, (iii) with missing data and several discrete and/or censored variables.

The approaches to be compared are: (i) **DPM-vs**: the proposed method, (ii) **DPM-cont**: the proposed method without accounting for discrete variables/censoring (i.e., assuming all continuous variables), (iii) **DPM**: the proposed method with variable selection turned off (i.e., a prior probability $\rho_j = 1$), (iv) **DPM-ind**: the approach of Kim et al. (2006) when all variables are continuous (i.e., assuming non-informative variables are independent of the rest), but modified to treat discrete variables/censoring and missing data when applicable just as the proposed approach, (v) **Mclust-vs**: the approach of Raftery and Dean (2006) implemented with the `clustvarsel` package in R. When there are missing data, Random Forest Imputation (Stekhoven and Bühlmann, 2012) implemented with the `missForest` package in R is used prior to application of `clustvarsel`. However, the Mclust-vs approach does not treat discrete variables differently and thus treats all variables as continuous and uncensored.

Each simulation case is described below. Figure 2 provides a graphical depiction of the problem for the first eight variables from the first of the 100 realizations of Case 2(c). Case 1 simulations resulted in very similar data patterns as well.

Case 1(a). $n = 150$, $p = 10$. The true model has $M = 3$ components with mixing proportions 0.5, 0.25, 0.25, respectively, and $\mathbf{y} \mid \phi$ is a multivariate normal with no censoring nor missing data. Only two variables $\mathbf{y}^{(1)} = [y_1, y_2]'$ are informative, with means of (2, 0), (0, 2), (-1.5, -1.5), unit variances, and correlations of 0.5, 0.5, -0.5 in each component, respectively. The non-informative variables $\mathbf{y}^{(2)} = [y_3, \dots, y_{10}]'$ are generated as *iid* $\mathcal{N}(0, 1)$.

Case 1(b). Same as the setup in 1(a) only the non-informative variables $\mathbf{y}^{(2)}$ are correlated with $\mathbf{y}^{(1)}$ through the relation $\mathbf{y}^{(2)} = \mathbf{B}\mathbf{y}^{(1)} + \boldsymbol{\varepsilon}$, where \mathbf{B} is a 8×2 matrix whose elements are distributed as *iid* $\mathcal{N}(0, 0.3)$, and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{Q}_{22}^{-1})$, with $\mathbf{Q}_{22} \sim \mathcal{W}(\mathbf{I}, 10)$.

Case 1(c). Same as in 1(b), but variables y_1, y_6 are discretized to the closest integer, variables y_2, y_9 are left censored at -1.4 ($\sim 8\%$ of the observations), and y_3, y_{10} are right censored at 1.4.

Case 1(d). Same as 1(c), but the even numbered y_j have $\sim 30\%$ of the observations MAR.

Case 2(a). $n = 300$, $p = 30$. The true model has $M = 3$ components with mixing proportions 0.5, 0.25, 0.25, respectively, $\mathbf{y} \mid \phi$ is a multivariate normal with no censoring nor missing data. Only four variables (y_1, y_2, y_3, y_4) are informative, with means of (0.6, 0, 1.2, 0), (0, 1.5, -0.6, 1.9), (-2, -2, 0, 0.6) and all variables with unit variance for each of the three components, respectively. All correlations among informative variables are equal to 0.5 in components 1 and 2, while component 3 has correlation matrix, $\boldsymbol{\Sigma}_{311}(i, j) = 0.5(-1)^{\|i+j\|} I_{\{i \neq j\}} + I_{\{i=j\}}$.

The non-informative variables $\mathbf{y}^{(2)} = [y_5, \dots, y_{30}]'$ are generated as *iid* $\mathcal{N}(0, 1)$.

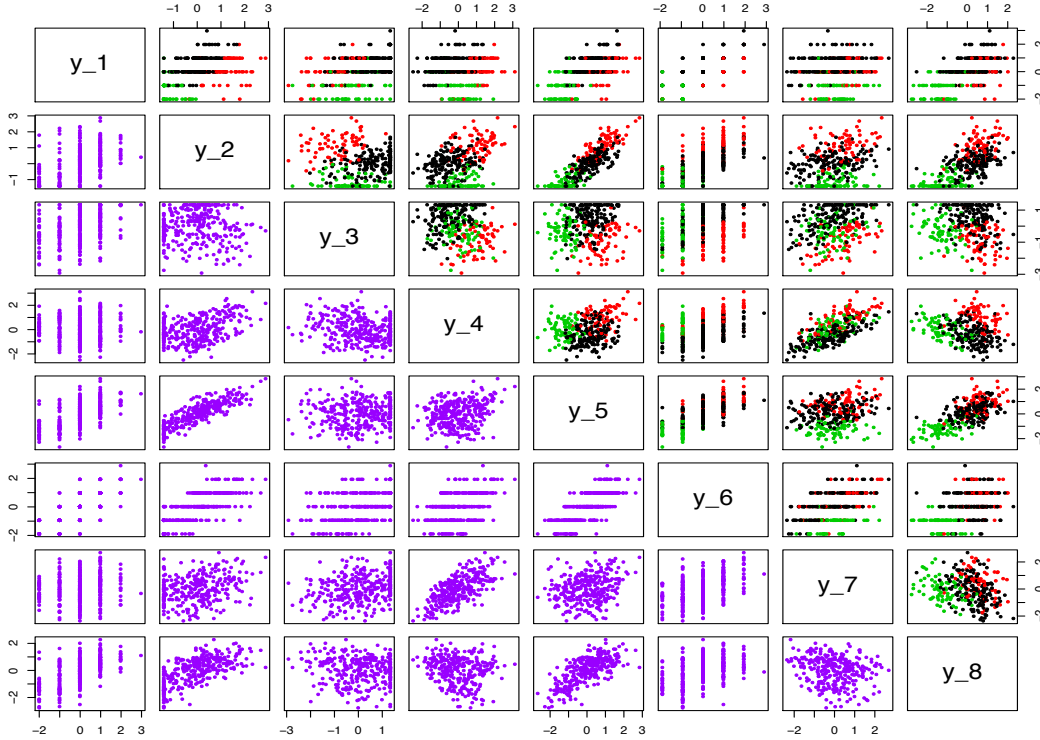
Case 2(b). Same as the setup in Case 2(a) only the non-informative variables $\mathbf{y}^{(2)}$ are correlated with $\mathbf{y}^{(1)}$ through the relation $\mathbf{y}^{(2)} = \mathbf{B}\mathbf{y}^{(1)} + \boldsymbol{\varepsilon}$, where \mathbf{B} is a 26×4 matrix whose elements are distributed as *iid* $\mathcal{N}(0, 0.3)$, and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{Q}_{22}^{-1})$, with $\mathbf{Q}_{22} \sim \mathcal{W}(\mathbf{I}, 30)$.

Case 2(c). Same setup as in Case 2(b), but now variables y_1, y_6, y_{11} are discretized to the closest integer, variables y_2, y_9, y_{10}, y_{11} are left censored at -1.4 ($\sim 8\%$ of the observations), and variables $y_3, y_{12}, y_{13}, y_{14}$ are right censored at 1.4.

Case 2(d). Same as Case 2(c), but the even numbered y_j have $\sim 30\%$ MAR.

For each of the eight simulation cases, 100 data sets were randomly generated and each of the five methods above was fit to each data set. The methods are compared on the basis of the following statistics: (i) Accuracy (Acc), calculated as the proportion of observations in the estimated clustering that are in the same group as they are in the true clustering, when put in the arrangement (relabeling) that best matches the true clusters. (ii) Fowlkes-Mallows index (FI) of $\hat{\phi}$ relative to the true clusters. (iii) Adjusted Rand index (ARI). (iv) The number of groups in the estimated clustering (M). (v) The model size, $p_1 = \sum_j \hat{\gamma}_j$. (vi) The proportion of variables correctly included/excluded from the model ($1/p \sum_j I_{\{\hat{\gamma}_j = \gamma_j\}}$) (PVC). (vii) The computation

Fig. 2: Pairwise scatter plots of the first eight variables for simulation Case 2(c).



time ($CompT$) in minutes (using 10,000 iterations for the Bayesian methods). These measures are summarized in the tables below by their mean (and stdev) over the 100 data sets.

The simulation results from Cases 1(a)-(d) are summarized in Table 1. The summary score for the *best* method for each summary is in bold font along with that for any other method that was not statistically different from the *best* method on the basis of the 100 trials (via an uncorrected paired t -test with a level of significance of 0.05). As might be expected, DPM-ind is one of the best methods on Case 1(a). However, it is not significantly better than DPM-vs or Mclust-vs on any of the metrics. All of the variable selection methods solidly outperform DPM, which had a difficult time finding more than a single cluster since it had to include all 10 variables. In Case 1(b) the assumptions of DPM-ind are being violated and it is unable to perform adequate variable selection. It must include far too many non-informative variables due to the correlation within $\mathbf{y}^{(2)}$ and between $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$. The clustering performance suffers as a result and like DPM, it has difficulty finding more than a single cluster. Mclust-vs performs well in this case, but DPM-vs (and DPM-cont) is significantly better on two of the metrics. In case 1(c) DPM-vs is now explicitly accounting for the discrete and left/right censored variables, while DPM-cont does not. When the discrete variables are incorrectly assumed to be continuous it tends to create separate clusters at some of the unique values of the discrete variables. This is because a very high likelihood can be obtained by normal distributions that are almost singular along the direction of the discrete variables. Thus, DPM-vs substantially outperforms DPM-cont and Mclust-vs, demonstrating the importance of explicitly treating the discrete nature of the data when clustering. Finally, Case 1(d) shows that the loss of 30% of the data for half of the variables (including an informative variable) does not degrade the performance of DPM-vs by much. Mclust-vs has *much* faster run-time than the Bayesian methods, however, when there are discrete variables and/or missing data, Mclust-vs did not perform nearly as well as DPM-vs.

Table 1: Simulation Case 1 Results.

<i>Method</i>	<i>Acc</i>	<i>FI</i>	<i>ARI</i>	<i>M</i>	<i>p</i> ₁	<i>PVC</i>	<i>CompT</i>
Case 1(a)							
DPM-vs*	0.91 (0.11)	0.86 (0.07)	0.78 (0.16)	2.9 (0.4)	2.0 (0.3)	0.99 (0.04)	296 (18)
DPM	0.37 (0.02)	0.58 (0.00)	0.00 (0.00)	1.0 (0.1)	10.0 (0.0)	0.20 (0.00)	341 (28)
DPM-ind	0.90 (0.13)	0.86 (0.08)	0.76 (0.20)	3.0 (0.6)	1.9 (0.4)	0.99 (0.04)	295 (23)
Mclust-vs	0.91 (0.10)	0.86 (0.07)	0.78 (0.15)	3.0 (0.4)	2.0 (0.2)	0.99 (0.06)	1 (0)
Case 1(b)							
DPM-vs*	0.89 (0.14)	0.85 (0.08)	0.76 (0.20)	3.0 (0.7)	1.9 (0.4)	0.98 (0.05)	267 (19)
DPM	0.37 (0.02)	0.58 (0.00)	0.00 (0.00)	1.0 (0.0)	10.0 (0.0)	0.20 (0.00)	292 (15)
DPM-ind	0.37 (0.02)	0.58 (0.00)	0.00 (0.00)	1.0 (0.0)	8.5 (0.8)	0.35 (0.08)	243 (18)
Mclust-vs	0.87 (0.12)	0.83 (0.10)	0.72 (0.19)	2.8 (0.4)	2.0 (0.1)	0.94 (0.12)	1 (0)
Case 1(c)							
DPM-vs	0.85 (0.14)	0.81 (0.08)	0.68 (0.20)	2.8 (0.6)	1.9 (0.4)	0.97 (0.07)	254 (18)
DPM-cont	0.62 (0.06)	0.52 (0.06)	0.31 (0.07)	4.9 (0.7)	2.0 (0.5)	0.89 (0.11)	296 (25)
DPM	0.37 (0.02)	0.58 (0.00)	0.00 (0.00)	1.0 (0.0)	10.0 (0.0)	0.20 (0.00)	285 (19)
DPM-ind	0.37 (0.02)	0.58 (0.00)	0.00 (0.00)	1.0 (0.0)	8.2 (1.0)	0.38 (0.10)	243 (24)
Mclust-vs	0.49 (0.10)	0.43 (0.09)	0.19 (0.12)	6.7 (1.5)	2.7 (0.8)	0.67 (0.14)	1 (0)
Case 1(d)							
DPM-vs	0.77 (0.19)	0.76 (0.10)	0.57 (0.25)	2.6 (0.8)	1.9 (0.6)	0.95 (0.09)	304 (21)
DPM-cont	0.62 (0.04)	0.52 (0.05)	0.31 (0.06)	4.8 (0.8)	1.9 (0.5)	0.89 (0.11)	355 (29)
DPM	0.37 (0.02)	0.58 (0.00)	0.00 (0.00)	1.0 (0.1)	10.0 (0.0)	0.20 (0.00)	343 (29)
DPM-ind	0.37 (0.02)	0.58 (0.00)	0.00 (0.00)	1.0 (0.0)	7.6 (1.3)	0.44 (0.13)	277 (40)
Mclust-vs	0.50 (0.10)	0.43 (0.08)	0.20 (0.11)	7.0 (1.1)	3.1 (0.9)	0.64 (0.14)	1 (0)
True	1.00	1.00	1.00	3.0	2.0	1.00	—

* DPM-cont is identical to DPM-vs for cases 1(a) and 1(b) and is therefore not listed separately.

Table 2: Simulation Case 2 Results.

<i>Method</i>	<i>Acc</i>	<i>FI</i>	<i>ARI</i>	<i>M</i>	<i>p</i> ₁	<i>PVC</i>	<i>CompT</i>
Case 2(a)							
DPM-vs*	0.89 (0.11)	0.85 (0.09)	0.74 (0.20)	3.0 (0.6)	3.8 (0.7)	0.99 (0.02)	544 (45)
DPM	0.50 (0.03)	0.61 (0.01)	0.00 (0.00)	1.4 (0.8)	30.0 (0.0)	0.13 (0.00)	1158 (120)
DPM-ind	0.90 (0.10)	0.85 (0.08)	0.75 (0.17)	3.1 (0.6)	3.9 (0.4)	1.00 (0.02)	543 (36)
Mclust-vs	0.91 (0.12)	0.88 (0.09)	0.78 (0.23)	2.9 (0.5)	4.0 (0.8)	0.98 (0.07)	8 (3)
Case 2(b)							
DPM-vs*	0.90 (0.10)	0.85 (0.08)	0.75 (0.17)	3.1 (0.6)	3.9 (0.5)	0.99 (0.03)	557 (27)
DPM	0.50 (0.03)	0.61 (0.01)	0.00 (0.00)	1.0 (0.0)	30.0 (0.0)	0.13 (0.00)	1009 (67)
DPM-ind	0.50 (0.03)	0.61 (0.01)	0.00 (0.00)	1.0 (0.0)	28.9 (0.2)	0.17 (0.01)	966 (71)
Mclust-vs	0.83 (0.14)	0.80 (0.12)	0.63 (0.24)	2.6 (0.5)	3.4 (0.8)	0.91 (0.10)	8 (3)
Case 2(c)							
DPM-vs	0.93 (0.03)	0.89 (0.05)	0.82 (0.07)	3.3 (0.6)	4.0 (0.2)	1.00 (0.02)	589 (47)
DPM-cont	0.62 (0.16)	0.57 (0.14)	0.34 (0.20)	4.4 (1.2)	3.4 (1.4)	0.87 (0.05)	597 (49)
DPM	0.50 (0.03)	0.61 (0.01)	0.00 (0.00)	1.0 (0.0)	30.0 (0.0)	0.13 (0.00)	1078 (54)
DPM-ind	0.50 (0.03)	0.61 (0.01)	0.00 (0.00)	1.0 (0.0)	28.8 (0.5)	0.17 (0.02)	1060 (101)
Mclust-vs	0.45 (0.06)	0.40 (0.04)	0.13 (0.06)	6.4 (1.3)	2.8 (1.0)	0.81 (0.04)	8 (24)
Case 2(d)							
DPM-vs	0.91 (0.08)	0.86 (0.08)	0.78 (0.14)	3.2 (0.6)	3.9 (0.4)	0.99 (0.03)	577 (54)
DPM-cont	0.61 (0.16)	0.55 (0.14)	0.32 (0.19)	4.3 (1.1)	3.2 (1.2)	0.87 (0.05)	581 (41)
DPM	0.50 (0.03)	0.61 (0.01)	0.00 (0.00)	1.3 (0.5)	30.0 (0.0)	0.13 (0.00)	1028 (85)
DPM-ind	0.50 (0.03)	0.61 (0.01)	0.00 (0.00)	1.0 (0.0)	28.2 (1.1)	0.19 (0.04)	958 (81)
Mclust-vs	0.46 (0.06)	0.41 (0.04)	0.14 (0.05)	6.7 (1.4)	3.0 (1.2)	0.81 (0.04)	6 (7)
True	1.00	1.00	1.00	3.0	4.0	1.00	—

The simulation results from Cases 2(a)-(d) are summarized in Table 2. A similar story line from Case 1 carries over into Case 2 where there are now $p = 30$ (four informative) variables and $n = 300$ observations. Namely, DPM-vs is not significantly different from DPM-ind or Mclust-vs on any of the summary measures for Case 2(a), with the exception of computation time. DPM-vs is the best method on all summary statistics (except *CompT*) by a sizeable margin on the remaining cases. While Mclust is much faster than DPM-vs, the cases of the most interest

Table 3: Posterior inclusion probabilities and sample means for the six most informative tests.

Variable	$\text{pr}(\gamma_j = 1)$	Cluster Means		
		1	2	3
Beery_standard	1.000	0.77	-1.02	-0.44
CompTsc_ol	1.000	0.46	-1.21	-0.01
WJ_Pass_Comprehen_z_Score	0.944	0.38	-1.26	0.30
Adaptive_Composite	0.889	0.44	-0.68	0.16
ach_abc_Attention	0.460	-0.18	0.21	0.04
ach_abc_AnxDep	0.427	-0.04	0.06	-0.01

in this paper are those with discrete variables, censoring and/or missing data (i.e., Cases 1(c), 1(d), 2(c), and 2(d)). In these cases, the additional computation time of DPM-vs might seem inconsequential relative to the enormous gain in accuracy. It is interesting that DPM-vs suffers far less from the missing values when moving from Case 2(c) to 2(d) than it did from Case 1(c) to 1(d). This is likely due to the fact that there are a larger number of observations to offset the additional complexity of a larger p . However, it is also likely that the additional (correlated) variables may help to reduce the posterior variance of the *imputed* values.

4. APPLICATION TO AUTISM AND RELATED DISORDERS

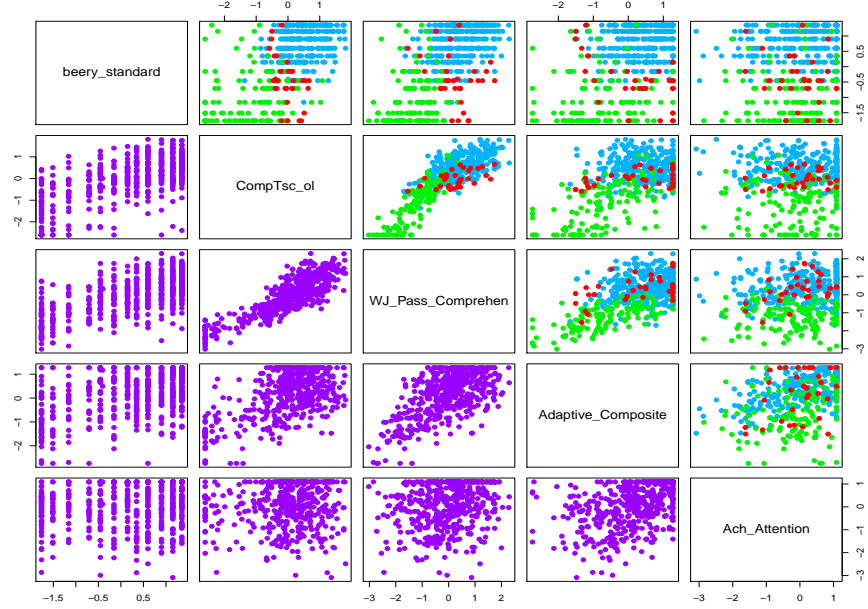
The cohort for this study consists of subjects falling in the criteria for “potential ASD” (PASD) on the basis of various combinations of developmental and psychiatric diagnoses obtained from comprehensive medical and educational records as described in Katusic et al. (2016). The population of individuals with PASD is important because this group represents the pool of patients with developmental/behavioral symptoms from which clinicians have to determine who has ASD and/or other disorders. Subjects 18 years of age or older were invited to participate in a face-to-face session to complete psychometrist-administered assessments of autism symptoms, cognition/intelligence, memory/learning, speech and language, adaptive functions, and maladaptive behavior. In addition, guardians were asked to complete several self-reported, validated questionnaires. The goal is to describe how the patients’ test scores separate them into different types of clinical presentation and which test scores are the most useful for this purpose. This falls in line with the new Research Domain Criteria (RDoC) philosophy that has gained traction in the field of mental health research. RDoC is a new research framework for studying mental disorders. It aims to integrate many levels of information (cognitive/self-report tests, imaging, genetics) to understand how all of these might be related to similar clinical presentations.

A total of 87 test scores measuring cognitive and/or behavioral characteristics were considered from a broad list of commonly used tests for assessing such disorders. Using expert judgment to include several commonly used aggregates in place of individual subtest scores, this list was reduced to 55 test score variables to be considered in the clustering procedure. Five of the 55 variables have fewer than 15 possible values and are treated here as discrete, ordinal variables. A majority (46) of the 55 variables also have a lower bound, which is attained by a significant portion of the individuals, and are treated as left censored. Five of the variables have an upper bound that is attained by many of the individuals and are thus treated as right censored. There are 479 observations (individuals) in the dataset, however, only 67 individuals have complete data, i.e., a complete case analysis would throw out 412 (86%) of the observations.

DPM-vs was applied to these data; four chains with random starting points were run in parallel for 85,000 iterations each, which took ~ 40 hours on a 2.2GHz processor. The first 10,000 iterations were discarded as burn-in. MCMC trace plots are provided in the Supplementary Material. All chains converged to the same distribution (aside from relabeling) and were thus combined.

Four of the tests (Beery standard, CompTsc_ol, WJ_Pass_Comprehen, and Adaptive Composite) had a high (> 0.88) posterior probability of being informative (Table 3). There is also

Fig. 3: Pairwise scatter plots of the standardized version of the five most informative variables in Table 3: with estimated cluster membership above the diagonal and the raw data below.



evidence that Ach_abc_Attention and Ach_abc_AnxDep are informative. The posterior samples were split on which of these two should be included in the model (they were only informative together for 0.1% of the MCMC samples). The next highest posterior inclusion probability for any of the remaining variables was 0.17 and the sum of the inclusion probabilities for all remaining variables was only 0.28. Thus, there is strong evidence to suggest that five variables are sufficient to inform the cluster membership.

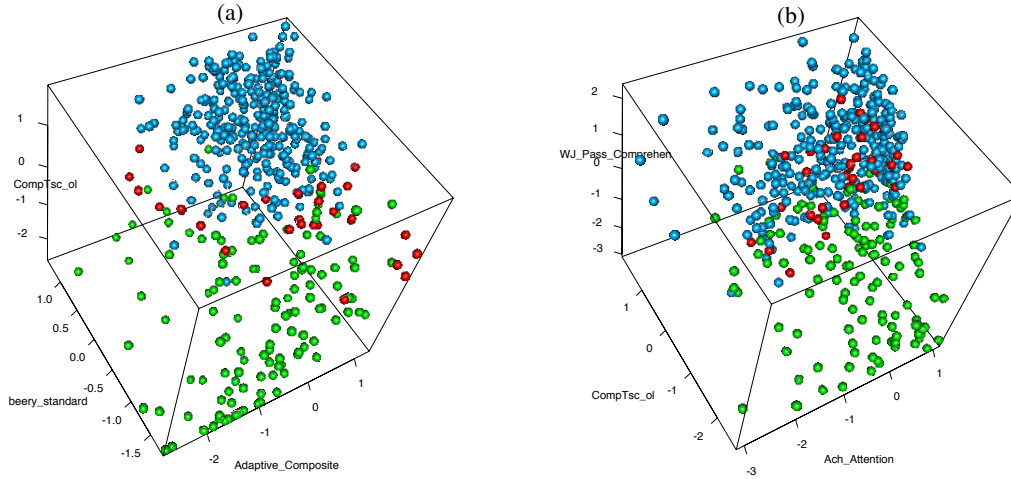
A majority (54%) of the posterior samples identified three components/clusters, with 0.12 and 0.25 posterior probability of two and four clusters, respectively. The calculation of $\hat{\phi}$ also resulted in three components. Figure 3 displays the estimated cluster membership via pairwise scatterplots of the five most informative variables on a standardized scale. Ach_abc_Attention has also been multiplied by minus one so that higher values imply better functioning for all tests. The corresponding mean vectors of the three main components are also provided in Table 3. It appears as though there are two groups that are very distinct (i.e., Clusters 1 and 2 are the “high” and “low” groups, respectively), but there is also a “middle” group (Cluster 3). Cluster 3 subjects generally have medium-to-high Adaptive_Composite, WJ_Pass_Comprehen, and Ach_abc_Attention scores, but low-to-medium Beery_standard and WJ_Pass_Comprehen.

Figure 4(a) provides a 3D plot of the clusters on the three most informative variables, highlighting the separation between Cluster 1 and Clusters 2 and 3. However, Clusters 2 and 3 are not well differentiated in this plot. Figure 4(b) shows a 3D scatter plot on the variables CompTsc_ol, WJ_Pass_Comprehen, and Ach_abc_Attention, to better illustrate some differentiation between Clusters 2 and 3. As these results are based on a bottom-up data-driven method, they may prove useful for determining imaging biomarkers that correspond better with cluster assignment than an arbitrary diagnosis provided by a physician. This will be the subject of future work.

5. CONCLUSIONS & FURTHER WORK

In this paper we developed a general approach to clustering via the Dirichlet process model that explicitly allows for (i) discrete and censored variables with a latent variable approach, (ii) missing data, (iii) correlation among informative and non-informative variables. The MCMC

Fig. 4: Three dimensional scatter plots of the tests on standardized scale: (a) The most informative three variables with estimated cluster membership. (b) Observations plotted on the variables `CompTsc_ol`, `WJ_Pass_Comprehen`, and `Ach_abc_Attention`, to better illustrate the separation of clusters 2 and 3.



computation proceeds via a split/merge type algorithm by integrating out the component parameters. This approach was shown to perform markedly better than other approaches on several simulated test cases. The approach was developed for moderate p in the range of $\sim 10 - 300$. The computation is $\mathcal{O}(p^3)$, which makes it ill suited for extremely large dimensions. However, it may be possible to use a sparse matrix approach, i.e., graphical model (Giudici and Green, 1999; Wong et al., 2003), within the proposed framework to alleviate this burden for very large p .

The approach was used to analyze the structure of test scores for individuals with potential ASD and it identified three primary clusters. Further, it determined that only five of the 55 variables were informative to assess the cluster membership of an observation. This could have a large impact for diagnosis of ASD as there are currently ~ 100 tests/subtest scores that could be used, and there is no universal standard. Further, the clustering results have served to generate hypotheses about what might show up in brain imaging to explain some of the differences between potential ASD patients. A follow-up study has been planned to investigate these possible connections.

REFERENCES

- Basu, S., and Chib, S. (2003), “Marginal likelihood and Bayes factors for Dirichlet process mixture models,” *Journal of the American Statistical Association*, 98(461), 224–235.
- Bhattacharya, A., and Dunson, D. B. (2012), “Simplex factor models for multivariate unordered categorical data,” *Journal of the American Statistical Association*, 107(497), 362–377.
- Bodnar, T., and Okhrin, Y. (2008), “Properties of the singular, inverse and generalized inverse partitioned Wishart distributions,” *Journal of Multivariate Analysis*, 99(10), 2389–2405.
- Chib, S., and Greenberg, E. (1998), “Analysis of multivariate probit models,” *Biometrika*, 85(2), 347–361.
- Chung, Y., and Dunson, D. B. (2012), “Nonparametric Bayes conditional distribution modeling with variable selection,” *Journal of the American Statistical Association*, .
- Dunson, D. B. (2000), “Bayesian latent variable models for clustered mixed outcomes,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(2), 355–366.
- Escobar, M. D., and West, M. (1995), “Bayesian density estimation and inference using mixtures,” *Journal of the American Statistical Association*, 90(430), 577–588.
- Ferguson, T. S. (1973), “A Bayesian Analysis of Some Nonparametric Problems,” *The Annals of Statistics*, 1(2), 209–230.
- Fop, M., Smart, K., and Murphy, T. B. (2015), “Variable Selection for Latent Class Analysis with Application to Low Back Pain Diagnosis,” *arXiv preprint arXiv:1512.03350*, .
- Fraley, C., and Raftery, A. (2002), “Model-based clustering, discriminant analysis, and density estimation,” *Journal of the American Statistical Association*, 97, 611–631.

- Friedman, J. H., and Meulman, J. J. (2004), "Clustering objects on subsets of attributes (with discussion)," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(4), 815–849.
- Fu, A. Q., Russell, S., Bray, S. J., Tavaré, S. et al. (2013), "Bayesian clustering of replicated time-course gene expression data with weak signals," *The Annals of Applied Statistics*, 7(3), 1334–1361.
- George, E., and McCulloch, R. (1993), "Variable selection via Gibbs sampling," *Journal of the American Statistical Association*, 88, 881–889.
- George, E., and McCulloch, R. (1997), "Approaches for Bayesian variable selection," *Statistica Sinica*, 7, 339–373.
- Giudici, P., and Green, P. (1999), "Decomposable graphical Gaussian model determination," *Biometrika*, 86(4), 785–801.
- Hennig, C., and Liao, T. F. (2013), "How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(3), 309–369.
- Hoff, P. D. (2006), "Model-based subspace clustering," *Bayesian Analysis*, 1(2), 321–344.
- Hunt, L., and Jorgensen, M. (2003), "Mixture model clustering for mixed data with missing information," *Computational Statistics & Data Analysis*, 41(3), 429–440.
- Imai, K., and van Dyk, D. A. (2005), "A Bayesian analysis of the multinomial probit model using marginal data augmentation," *Journal of Econometrics*, 124(2), 311–334.
- Ishwaran, H., and James, L. F. (2011), "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, .
- Jain, S., and Neal, R. M. (2004), "A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model," *Journal of Computational and Graphical Statistics*, 13(1), 158–182.
- Katusic, S. K., Myers, S., Colligan, R. C., Voigt, R., Stoeckel, R. E., Port, J. D., Croarkin, P. E., and Weaver, A. (2016), "Developmental brain dysfunction-related disorders and potential autism spectrum disorder (PASD) among children and adolescents - population based 1976-2000 birth cohort," *Lancet Neurology (in review)*, .
- Kim, S., Tadesse, M. G., and Vannucci, M. (2006), "Variable selection in clustering via Dirichlet process mixture models," *Biometrika*, 93(4), 877–893.
- Lesaffre, E., and Molenberghs, G. (1991), "Multivariate probit analysis: a neglected procedure in medical statistics," *Statistics in Medicine*, 10(9), 1391–1403.
- Lid Hjort, N., Holmes, C., Müller, P., and Walker, S. G. (2010), *Bayesian Nonparametrics*, New York, NY: Cambridge University Press.
- Lin, Y., and Zhang, H. (2006), "Component Selection and Smoothing in Smoothing Spline Analysis of Variance Models," *Annals of Statistics*, 34, 2272–2297.
- Liu, J., Zhang, J., Palumbo, M., and Lawrence, C. (2003), Bayesian clustering with variable and transformation selections, in *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, eds. J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, Oxford University Press, USA, pp. 249–275.
- MacEachern, S. N., and Müller, P. (1998), "Estimating mixture of Dirichlet process models," *Journal of Computational and Graphical Statistics*, 7(2), 223–238.
- Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2009), "Variable selection for clustering with Gaussian mixture models," *Biometrics*, 65(3), 701–709.
- McCulloch, R. E., Polson, N. G., and Rossi, P. E. (2000), "A Bayesian analysis of the multinomial probit model with fully identified parameters," *Journal of Econometrics*, 99(1), 173–193.
- Murray, J. S., and Reiter, J. P. (2016), "Multiple Imputation of Missing Categorical and Continuous Values via Bayesian Mixture Models with Local Dependence," *arXiv preprint arXiv:1410.0438*, .
- Muthén, B. (1983), "Latent variable structural equation modeling with categorical data," *Journal of Econometrics*, 22(1), 43–65.
- Neal, R. M. (2000), "Markov chain sampling methods for Dirichlet process mixture models," *Journal of computational and graphical statistics*, 9(2), 249–265.
- Pan, W., and Shen, X. (2007), "Penalized model-based clustering with application to variable selection," *Journal of Machine Learning Research*, 8(May), 1145–1164.
- Quintana, F. A., and Iglesias, P. L. (2003), "Bayesian clustering and product partition models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2), 557–574.
- Raftery, A. E., and Dean, N. (2006), "Variable selection for model-based clustering," *Journal of the American Statistical Association*, 101(473), 168–178.
- Reich, B., Storlie, C., and Bondell, H. (2009), "Variable Selection in Bayesian Smoothing Spline ANOVA Models: Application to Deterministic Computer Codes," *Technometrics*, 51, 110–120.
- Richardson, S., and Green, P. J. (1997), "On Bayesian analysis of mixtures with an unknown number of components (with discussion)," *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4), 731–792.
- Rubin, D. B. (1976), "Inference and missing data," *Biometrika*, 63(3), 581–592.
- Rue, H., and Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, Boca Raton, FL: Chapman & Hall/CRC.
- Sethuraman, J. (1994), "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639–650.

- Stekhoven, D. J., and Bühlmann, P. (2012), “MissForest: non-parametric missing value imputation for mixed-type data,” *Bioinformatics*, 28(1), 112–118.
- Stephens, M. (2000), “Dealing with label switching in mixture models,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4), 795–809.
- Storlie, C. B., Lane, W. A., Ryan, E. M., Gattiker, J. R., and Higdon, D. M. (2015), “Calibration of computational models with categorical parameters and correlated outputs via Bayesian smoothing spline ANOVA,” *Journal of the American Statistical Association*, 110(509), 68–82.
- Storlie, C., Therneau, T., Carter, R., Chia, N., Boughey, J., Bergquist, J., and Romero-Brufau, S. (2017), “Prediction and Inference with Missing Data in Patient Alert Systems,” *Journal of the American Statistical Association (in review)*, .
- Tadesse, M. G., Sha, N., and Vannucci, M. (2005), “Bayesian variable selection in clustering high-dimensional data,” *Journal of the American Statistical Association*, 100(470), 602–617.
- Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006), “Hierarchical Dirichlet processes,” *Journal of the American Statistical Association*, 101, 1566–1581.
- Tibshirani, R. (1996), “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society B*, 58, 267–288.
- Wang, S., and Zhu, J. (2008), “Variable Selection for Model-Based High-Dimensional Clustering and Its Application to Microarray Data,” *Biometrics*, 64(2), 440–448.
- Witten, D. M., and Tibshirani, R. (2012), “A framework for feature selection in clustering,” *Journal of the American Statistical Association*, .
- Wong, F., Carter, C. K., and Kohn, R. (2003), “Efficient estimation of covariance selection models,” *Biometrika*, 90(4), 809–830.
- Xie, B., Pan, W., and Shen, X. (2008), “Variable Selection in Penalized Model-Based Clustering Via Regularization on Grouped Parameters,” *Biometrics*, 64(3), 921–930.
- Zhang, X., Boscardin, W. J., and Belin, T. R. (2008), “Bayesian analysis of multivariate nominal measures using multivariate multinomial probit models,” *Computational statistics & data analysis*, 52(7), 3697–3708.
- Zou, H., and Hastie, T. (2005), “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.

Supplementary Material: “Clustering and Variable Selection in the Presence of Mixed Variable Types and Missing Data”

A. COGNITIVE/BEHAVIORAL TESTS DESCRIPTIONS

Table S.1: Test and self-report form descriptions for the 55 tests used in the analysis in the main paper.

Variable Name	Test/Form	Description
General_Adaptive_Composite	Adaptive Behavior Assessment System (ABAS-II) overall adaptive functioning composite score	Includes all 9 skill areas in the 3 rows below plus Work (when applicable).
Conceptual_Composite_Score	ABAS-II Conceptual Composite domain	Includes Communication, Functional Academics, and Self-Direction skill areas.
Social_Composite_Score	ABAS-II Social Composite domain	Includes Leisure and Social skill areas.
Practical_Composite_Score	ABAS-II Practical Composite domain	Includes Community Use, Home Living, Health and Safety, and Self-Care skill areas.
ABC_Irritability_raw	Aberrant Behavior Checklist, Irritability scale	ABC is a maladaptive behavior rating scale. Higher scores are worse on all of the ABC subscales. Normed for individuals with significant developmental disabilities requiring special education, so does not capture mild issues in the general population.
ABC_Lethargy_raw	ABC Lethargy/Social Withdrawal scale.	Items reflect underactivity or listlessness, social withdrawal (seeking isolation, unresponsiveness to social interactions, etc.).
ABC_Stereotype_raw	ABC Stereotype scale.	Stereotype scale measures compulsions and repetitive stereotyped behaviors.
ABC_Hyperactivity_raw	ABC Hyperactivity scale	Includes ADHD symptoms (inattention, distractibility, impulsivity, hyperactivity) and also noncompliance/oppositional behavior.
ABC_Inappropriate_Speech_raw	ABC Inappropriate Speech scale	Only 4 items, which capture the following aspects of speech: excessive, repetitive (2 items), self-directed and loud.
COM_RSI_Total	Autism Diagnostic Observation Schedule (ADOS), Communication + Reciprocal Social Interaction score	The Communication + Reciprocal Social Interaction total score is used to determine ADOS-2 classification (autism, autism spectrum, or non-spectrum) based on cutoff scores. This score corresponds with the DSM-5 Social Communication criteria (and does not include the restricted and repetitive behavior aspect).
SBRI_Total	Autism Diagnostic Observation Schedule (ADOS), Stereotyped Behaviors and Restricted Interests score	This score includes unusual sensory interests/behaviors, stereotyped mannerisms, circumscribed or unusual interests, and compulsions/rituals. The SBRI score is not used in the ADOS-2 diagnostic algorithm but informs determination of whether DSM-5 restricted interests/activities and repetitive behavior criteria are met. Note: Sometimes these behaviors are not exhibited during testing yet are prominent at home and in the community - only observed behaviors can be scored. So, it may underestimate RRB.
Beery_standard	Beery-Buktenica Developmental Test of Visual-Motor Integration (Beery VMI)	Assesses the extent to which individuals can integrate their visual and motor abilities (degree to which visual perception and finger-hand movements are well coordinated). Important role in the development of handwriting and other skills.

Inhibit.T	BRIEF Inhibition	Ability to inhibit impulsive responses (resist impulses, stop one's own behavior at the appropriate time).
Shift.T	BRIEF Task Shift	Ability to adjust to changes in routine or task demands. Key aspects of shifting include the ability to make transitions, tolerate change, problem-solve flexibly, switch or alternate attention, and change focus from one mindset or topic to another.
Emotional_Control.T	BRIEF Emotional Control	Measures the impact of executive function problems on emotional expression and assesses a child's ability to modulate or control his or her emotional responses.
Self_Monitor.T	BRIEF Self-monitoring	Self-monitoring or interpersonal awareness (whether a child keeps track of the effect that his or her behavior has on others).
Initiate.T	BRIEF Task Initiation	Ability to begin a task or activity and to independently generate ideas, responses, or problem-solving strategies.
Working_Memory.T	BRIEF Working Memory	Ability to hold information in mind for the purpose of completing a task, encoding information, or generating goals, plans, and sequential steps to achieving goals.
Plan_Organize.T	BRIEF Task Organization	Ability to manage current and future-oriented task demands (plan and organize problem solving approaches).
Task_Monitor.T	BRIEF Task Monitoring	Task-oriented monitoring or work-checking habits (whether a child assesses his or her own performance during or shortly after finishing a task to ensure accuracy or appropriate attainment of a goal).
Org_of_Materials.T	BRIEF Organization of Materials	Ability to organize environment and materials - orderliness of work, play, and storage spaces (e.g., desks, lockers, backpacks, and bedrooms).
BRI.T	BRIEF Behavioral Regulation Index (BRI)	Summary capturing ability to shift cognitive set and modulate emotions and behavior via appropriate inhibitory control. Includes Inhibit, Shift, and Emotional Control subscales.
MI.T	BRIEF Metacognition Index (MI)	Summary capturing ability to initiate, plan, organize, self-monitor, and sustain working memory - relates directly to a child's ability to actively problem solve in a variety of contexts. Includes Initiate, Working Memory, Plan/Organize, Organization of Materials, and Monitor subscales.
GEC.T	BRIEF Global Executive Composite (GEC)	Overall index of executive function; incorporates all of the BRIEF clinical scales.
Cars.Total	Childhood Autism Rating Scales (CARS2-ST and CARS2-HF)	Structured interview and observation tool. Scores are raw scores (T-scores and percentiles among population of individuals with ASD are available). A measure of overall severity of ASD-related symptoms based on 15 items. Ratings are based not only on frequency of the behavior in question, but also on its intensity, atypicality, and duration.
Tsc.lc	Oral and Written Language Scales (OWLS-II), Listening Comprehension (LC) subtest	Measures oral language reception, or understanding of spoken language. Examiner orally presents increasingly difficult words, phrases, and sentences; patient responds by pointing to or stating which of four picture choices is correct.

Tsc.oe	Oral and Written Language Scales (OWLS-II), Oral Expression (OE) subtest	Measures oral language expression, or use of spoken language. Examiner presents a verbal prompt along with a picture and patient must respond orally to the prompt with increasingly difficult language.
CompTsc.ol	Oral and Written Language Scales (OWLS-II), Oral Language Composite	Represents an overall level of oral language functioning. Derived from the Listening Comprehension and Oral Expression scales.
scq_raw_total	Social Communication Questionnaire (Lifetime Version)	40-item yes/no questionnaire; many items focus on the presence of symptoms during the period between the individual's 4th and 5th birthdays. Scores are raw scores (no standardized scores available). Designed to assess for qualitative impairments in reciprocal social interaction and communication, as well as restricted, repetitive, and stereotyped behavior
T_RRB	Social Responsiveness Scale (SRS) Restricted Interests and Repetitive Behavior T-score	Items assess restricted range of interests and activities, inflexibility, unusual sensory interests, perseveration on topics, atypicality (bizarre behavior, being regarded as odd by peers) as well as motor stereotypy.
T_Score	Social Responsiveness Scale (SRS) Total T-score	Reflects the sum of responses to all 65 SRS questions (including the SCI and RRB subscales). Serves as an index of reciprocal social behavior across typical development, ASD, and other disorders. A good single number rating of severity of ASD symptoms.
T_SCI	Social Responsiveness Scale (SRS) Social Communication and Interaction (SCI) T-score	Reflects 4 subscales: Social Awareness, Social Cognition, Social Communication, Social Motivation
wasi_iq_composite	Wechsler Abbreviated Scale of Intelligence (WASI-II)	IQ composite score based on Vocabulary, Similarities, Block Design, Matrix Reasoning subtests.
WJ_Basic_Read_Skills_z_Score	Woodcock-Johnson Test of Achievement, Basic Reading cluster.	Measures sight vocabulary and the ability to apply phonic and structural analysis skills. Combination of Letter-Word Identification and Word Attack.
WJ_Pass_Comprehen_z_Score	Woodcock-Johnson Test of Achievement,	Reading comprehension. Measures understanding of written text. The majority of items require a student to supply a missing word to sentences and then paragraphs of increasing complexity.
WJ_Word_Attack_z_Score	Woodcock-Johnson Test of Achievement, Word Attack subtest	Measures ability to apply phonic/decoding skills to unfamiliar words. The majority of items require students to pronounce non-sense words of increasing complexity.
wraml_Verbal_Memory_Index_Sum	Wide Range Assessment of Memory and Learning (WRAML-2)	Measures ability to learn and recall both meaningful verbal information and relatively rote verbal information. Derived from the sum of the Story Memory and Verbal Learning subtests.
wrat_spelling_standard	Wide Range Achievement Test (WRAT4), Spelling subtest	Measures ability to identify sounds and transfer them into written form from dictated words. Standard spelling test - word is stated, used in a sentence, and repeated and patient writes it.
wrat_math_standard	Wide Range Achievement Test (WRAT4), Math Computation subtest	Measures ability to count, identify numbers, solve simple oral math problems, and calculate written math problems. Problems are presented in a range of domains, including arithmetic, algebra, geometry, and advanced operations.

ach.abc.AnxDep	Achenbach Assessment of Empirically Based Assessment (ASEBA) - Adult Behavior Checklist (ABCL) or Adult Self-Report (ASR) - Anxious/Depressed scale	Measures behaviors such as nervousness, worrying, fearfulness, loneliness, sadness, feeling worthless, feeling too guilty, feeling persecuted, lacking self-confidence.
ach.abc.Withdrawn	ASEBA - Adult Behavior Checklist (ABCL) or Adult Self-Report (ASR) - Withdrawn scale	Measures behaviors such as poor relationships, not getting along with others, preferring to be alone, anhedonia, being secretive.
ach.abc.Somatic	ASEBA - Adult Behavior Checklist (ABCL) or Adult Self-Report (ASR) - Somatic Complaints scale	Measures complaints of discomfort or illness.
ach.abc.Thought	ASEBA - Adult Behavior Checklist (ABCL) or Adult Self-Report (ASR) - Thought Problems scale	Measures symptoms such as hallucinations, obsessions, compulsions, strange thoughts and behaviors, self-harm, and suicide attempts.
ach.abc.Attention	ASEBA - Adult Behavior Checklist (ABCL) or Adult Self-Report (ASR) - Attention Problems scale	Measures attention problems, forgetfulness, daydreaming, failing to finish things, avoiding work, disorganization, lateness, difficulty planning and prioritizing.
ach.abc.Agressive	ASEBA - Adult Behavior Checklist (ABCL) or Adult Self-Report (ASR) - Aggressive Behavior scale	Measures behaviors such as meanness, arguing, threatening, blaming others, fighting, temper outbursts, screaming, sulking.
ach.abc.RuleBreak	ASEBA - Adult Behavior Checklist (ABCL) or Adult Self-Report (ASR) - Rule-Breaking Behavior scale	Measures behaviors such as irresponsibility, substance abuse, lacking feelings of guilt, lying or cheating, stealing, difficulty keeping a job.
ach.abc.Intrusive	ASEBA - Adult Behavior Checklist (ABCL) or Adult Self-Report (ASR) - Intrusive scale	Measures behaviors such as bragging, showing off, attention-seeking, being boisterous, teasing.

B. MARGINALIZED LIKELIHOOD

The derivation of Result 3 in the main paper is provided below. Let the component parameters be denoted as $\theta = \{\boldsymbol{\mu}_{11}, \dots, \boldsymbol{\mu}_{M1}, \boldsymbol{\Sigma}_{111}, \dots, \boldsymbol{\Sigma}_{M11}, \mathbf{b}_2, \mathbf{Q}_{21}, \mathbf{Q}_{22}\}$. We wish to obtain a closed form result for,

$$f(\mathbf{Z} \mid \boldsymbol{\gamma}, \boldsymbol{\phi}) = \int f(\mathbf{Z} \mid \boldsymbol{\gamma}, \boldsymbol{\phi}, \theta) f(\theta \mid \boldsymbol{\gamma}) d\theta.$$

However, after a little bit of algebra we have,

$$\begin{aligned} f(\mathbf{Z} \mid \boldsymbol{\gamma}, \boldsymbol{\phi}, \theta) &= \prod_{m=1}^M \prod_{\{i: \phi_i = m\}} \frac{1}{(2\pi)^{p/2}} |\boldsymbol{\Sigma}_m|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{z}_i - \boldsymbol{\mu}_m)' \boldsymbol{\Sigma}_m^{-1} (\mathbf{z}_i - \boldsymbol{\mu}_m) \right\} \\ &= \left[\prod_{m=1}^M A_m \right] B, \end{aligned}$$

where $M = \max_i \{\phi_i\}$, and

$$\begin{aligned} A_m &= (2\pi)^{-\frac{n_m p_1}{2}} |\boldsymbol{\Sigma}_{m11}|^{-\frac{n_m}{2}} \exp \left\{ -\frac{1}{2} \sum_{i: \phi_i = m} (\mathbf{z}_i^{(1)} - \boldsymbol{\mu}_{m1})' \boldsymbol{\Sigma}_{m11}^{-1} (\mathbf{z}_i^{(1)} - \boldsymbol{\mu}_{m1}) \right\}, \text{ and} \\ B &= (2\pi)^{-\frac{n p_2}{2}} |\mathbf{Q}_{22}|^{\frac{n}{2}} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n \left[\mathbf{z}_i^{(2)'} \mathbf{Q}_{22} \mathbf{z}_i^{(2)} - 2 \mathbf{z}_i^{(2)'} (\mathbf{b}_2 - \mathbf{Q}_{21} \mathbf{z}_i^{(1)}) + (\mathbf{b}_2 - \mathbf{Q}_{21} \mathbf{z}_i^{(1)})' \mathbf{Q}_{22} (\mathbf{b}_2 - \mathbf{Q}_{21} \mathbf{z}_i^{(1)}) \right] \right\}. \end{aligned}$$

Combining this with the prior independence of $(\boldsymbol{\mu}_{m1}, \boldsymbol{\Sigma}_{m11})$, $m = 1, \dots$ and $(\mathbf{b}_2, \mathbf{Q}_{21}, \mathbf{Q}_{22})$ we have,

$$\int f(\mathbf{Z} \mid \gamma, \phi, \theta) f(\theta \mid \gamma) d\theta = \left[\prod_{m=1}^M \int A_m f(\boldsymbol{\mu}_{m1}, \boldsymbol{\Sigma}_{m11}) d(\boldsymbol{\mu}_{m1}, \boldsymbol{\Sigma}_{m11}) \right] \int B f(\mathbf{b}_2, \mathbf{Q}_{21}, \mathbf{Q}_{22}) d(\mathbf{b}_2, \mathbf{Q}_{21}, \mathbf{Q}_{22}). \quad (\text{S.1})$$

Now

$$f(\boldsymbol{\mu}_{m1}, \boldsymbol{\Sigma}_{m11}) = f(\boldsymbol{\mu}_{m1} \mid \boldsymbol{\Sigma}_{m11}) f(\boldsymbol{\Sigma}_{m11}),$$

with

$$f(\boldsymbol{\mu}_{m1} \mid \boldsymbol{\Sigma}_{m11}) = (2\pi)^{-\frac{p_1}{2}} \left| \frac{1}{\lambda} \boldsymbol{\Sigma}_{m11} \right|^{-\frac{1}{2}} \exp \left\{ -\frac{\lambda}{2} \boldsymbol{\mu}_{m1}' \boldsymbol{\Sigma}_{m11}^{-1} \boldsymbol{\mu}_{m1} \right\}$$

$$f(\boldsymbol{\Sigma}_{m11}) = \frac{|\boldsymbol{\Psi}_{11}|^{-\frac{\eta-p_2}{2}} |\boldsymbol{\Sigma}_{m11}|^{-\frac{\eta-p_2+p_1+1}{2}}}{2^{\frac{(\eta-p_2)p_1}{2}} \Gamma_{p_1}(\frac{\eta-p_2}{2})} \exp \left\{ -\frac{1}{2} \text{tr}(\boldsymbol{\Psi}_{11} \boldsymbol{\Sigma}_{m11}^{-1}) \right\}.$$

After some tedious algebra,

$$A_m f(\boldsymbol{\mu}_{m1}, \boldsymbol{\Sigma}_{m11}) = A_m^{(1)} A_m^{(2)} A_m^{(3)},$$

with

$$A_m^{(1)} = \pi^{-\frac{n_m p_1}{2}} \left(\frac{\lambda}{n_m + \lambda} \right)^{\frac{p_1}{2}} \frac{|\boldsymbol{\Psi}_{11}|^{\frac{\eta-p_2}{2}} \Gamma_{p_1}(\frac{n_m+\eta-p_2}{2})}{|\mathbf{V}_{m11}|^{\frac{\eta-p_2}{2}} \Gamma_{p_1}(\frac{\eta-p_2}{2})},$$

$$A_m^{(2)} = 2\pi^{-\frac{p_1}{2}} \left| \frac{1}{n_m + \lambda} \boldsymbol{\Sigma}_{m11} \right|^{-\frac{1}{2}} \exp \left\{ -\frac{n_m + \lambda}{2} \left(\boldsymbol{\mu}_{m1} - \frac{n_m}{n_m + \lambda} \bar{\mathbf{z}}_{m1} \right)' \boldsymbol{\Sigma}_{m11}^{-1} \left(\boldsymbol{\mu}_{m1} - \frac{n_m}{n_m + \lambda} \bar{\mathbf{z}}_{m1} \right) \right\}$$

$$A_m^{(3)} = \frac{|\mathbf{V}_{m11}|^{\frac{n_m+\eta-p_2}{2}} |\boldsymbol{\Sigma}_{m11}|^{-\frac{n_m+\eta-p_2+p_1+1}{2}}}{2^{\frac{(n_m+\eta-p_2)p_1}{2}} \Gamma_{p_1}(\frac{n_m+\eta-p_2}{2})} \exp \left\{ -\frac{1}{2} \text{tr}(\mathbf{V}_{m11} \boldsymbol{\Sigma}_{m11}^{-1}) \right\},$$

where \mathbf{V}_{m11} and $\bar{\mathbf{z}}_{m1}$ are as defined in Result 3 of the main paper. As a function of $\boldsymbol{\mu}_{m1}$, we recognize $A_m^{(2)}$ to be the multivariate normal density with mean $\frac{n_m}{n_m + \lambda} \bar{\mathbf{z}}_{m1}$ and covariance $\frac{1}{n_m + \lambda} \boldsymbol{\Sigma}_{m11}$. Also, as a function of $\boldsymbol{\Sigma}_{m11}$ we recognize $A_m^{(3)}$ to be the density of an inverse-Wishart distribution with parameters $\eta^* = n_m + \eta - p_2$ and $\boldsymbol{\Psi}^* = \mathbf{V}_{m11}$. Thus,

$$\int A_m f(\boldsymbol{\mu}_{m1}, \boldsymbol{\Sigma}_{m11}) d(\boldsymbol{\mu}_{m1}, \boldsymbol{\Sigma}_{m11}) = \pi^{-\frac{n p_1}{2}} \prod_{m=1}^M \left[\left(\frac{\lambda}{n_m + \lambda} \right)^{\frac{p_1}{2}} \frac{|\boldsymbol{\Psi}_{11}|^{\frac{\eta-p_2}{2}} \Gamma_{p_1}(\frac{n_m+\eta-p_2}{2})}{|\mathbf{V}_{m11}|^{\frac{n_m+\eta-p_2}{2}} \Gamma_{p_1}(\frac{\eta-p_2}{2})} \right]. \quad (\text{S.2})$$

Now the prior distribution corresponding to the second term in (S.1) is

$$f(\mathbf{b}_2, \mathbf{Q}_{21}, \mathbf{Q}_{22}) = f(\mathbf{b}_2 \mid \mathbf{Q}_{22}) f(\mathbf{Q}_{21} \mid \mathbf{Q}_{22}) f(\mathbf{Q}_{22}),$$

where

$$f(\mathbf{b}_2 \mid \mathbf{Q}_{22}) = (2\pi)^{-\frac{p_2}{2}} \left| \frac{1}{\lambda} \mathbf{Q}_{22} \right|^{-\frac{1}{2}} \exp \left\{ -\frac{\lambda}{2} \mathbf{b}_2' \mathbf{Q}_{22}^{-1} \mathbf{b}_2 \right\}$$

$$f(\mathbf{Q}_{21} \mid \mathbf{Q}_{22}) = (2\pi)^{-\frac{p_1 p_2}{2}} |\boldsymbol{\Psi}_{11}|^{\frac{p_2}{2}} |\mathbf{Q}_{22}|^{-\frac{p_1}{2}} \exp \left\{ -\frac{1}{2} \text{tr} \left[\boldsymbol{\Psi}_{11} (\mathbf{Q}_{21} + \mathbf{Q}_{22} \boldsymbol{\Psi}_{21} \boldsymbol{\Psi}_{11}^{-1})' \mathbf{Q}_{22}^{-1} (\mathbf{Q}_{21} + \mathbf{Q}_{22} \boldsymbol{\Psi}_{21} \boldsymbol{\Psi}_{11}^{-1}) \right] \right\}$$

$$f(\mathbf{Q}_{22}) = \frac{|\boldsymbol{\Psi}_{22|1}|^{\frac{\eta}{2}} |\mathbf{Q}_{22}|^{\frac{\eta+p_2+1}{2}}}{2^{\frac{\eta p_2}{2}} \Gamma_{p_2}(\frac{\eta}{2})} \exp \left\{ -\frac{1}{2} \text{tr}(\boldsymbol{\Psi}_{22|1} \mathbf{Q}_{22}) \right\},$$

where $\Psi_{22|1} = \Psi_{22} - \Psi_{21}\Psi_{11}\Psi_{12}$. After some more tedious algebra,

$$Bf(\mathbf{b}_2, \mathbf{Q}_{21}, \mathbf{Q}_{22}) = B^{(1)}B^{(2)}B^{(3)}B^{(4)},$$

with

$$\begin{aligned} B^{(1)} &= \pi^{-\frac{np_2}{2}} \left(\frac{\lambda}{n+\lambda} \right)^{\frac{p_2}{2}} \frac{|\Psi_{11}|^{\frac{\eta-p_2}{2}} |\Psi_{22|1}|^{\frac{\eta}{2}} \Gamma_{p_2}(\frac{n+\eta}{2})}{|V_{11}|^{\frac{p_2}{2}} |V_{22|1}|^{\frac{n+\eta}{2}} \Gamma_{p_2}(\frac{\eta}{2})}, \\ B^{(2)} &= 2\pi^{-\frac{p_2}{2}} \left| \frac{1}{n+\lambda} \mathbf{Q}_{22} \right|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \left[\mathbf{b}_2' \mathbf{Q}_{22}^* \mathbf{b}_2 - 2\mathbf{b}_2' \mathbf{b}^* + \mathbf{b}^{*'} \mathbf{Q}^{*-1} \mathbf{b}^* \right] \right\} \\ B^{(3)} &= (2\pi)^{-\frac{p_1 p_2}{2}} |V_{11}|^{-\frac{p_2}{2}} |\mathbf{Q}_{22}|^{-\frac{p_1}{2}} \exp \left\{ -\frac{1}{2} \text{tr} \left[V_{11} (\mathbf{Q}_{21} + \mathbf{Q}_{22} V_{21} V_{11}^{-1})' \mathbf{Q}_{22}^{-1} (\mathbf{Q}_{21} + \mathbf{Q}_{22} V_{21} V_{11}^{-1}) \right] \right\} \\ B^{(4)} &= \frac{|V_{22|1}|^{-\frac{n+\eta}{2}} |\mathbf{Q}_{22}|^{-\frac{n+\eta+p_2+1}{2}}}{2^{\frac{(n+\eta)p_2}{2}} \Gamma_{p_2}(\frac{n+\eta}{2})} \exp \left\{ -\frac{1}{2} \text{tr} (V_{22|1} \mathbf{Q}_{22}) \right\}, \end{aligned}$$

where $\mathbf{b}^* = n(\bar{\mathbf{y}}_2 + \mathbf{Q}_{22}^{-1} \mathbf{Q}_{21} \bar{\mathbf{y}}_1)$ and $\mathbf{Q}^* = (n+\lambda)\mathbf{Q}_{22}^{-1}$, and V_{11} , $V_{22|1}$ are as defined in Result 3 of the main paper.

As a function of \mathbf{b}_2 , we recognize $B^{(2)}$ to be the multivariate normal density in canonical form with precision \mathbf{Q}^* and mean $\mathbf{Q}^{*-1} \mathbf{b}^*$. Also, as a function of \mathbf{Q}_{21} we recognize $B^{(3)}$ to be the density of a $\mathcal{MN}(M^*, U^*, V^*)$ with parameters $M^* = -\mathbf{Q}_{22} V_{21} V_{11}^{-1}$, $U^* = \mathbf{Q}_{22}$, and $V^* = V_{11}^{-1}$. Finally, we can recognize $B^{(4)}$ to be the density of \mathbf{Q}_{22} : a Wishart distribution with parameters $\eta^* = n + \eta$ and $\Psi^* = V_{22|1}$. Thus,

$$\int Bf(\mathbf{b}_2, \mathbf{Q}_{21}, \mathbf{Q}_{22}) d(\mathbf{b}_2, \mathbf{Q}_{21}, \mathbf{Q}_{22}) = \pi^{-\frac{np_2}{2}} \left(\frac{\lambda}{n+\lambda} \right)^{\frac{p_2}{2}} \frac{|\Psi_{11}|^{\frac{\eta-p_2}{2}} |\Psi_{22|1}|^{\frac{\eta}{2}} \Gamma_{p_2}(\frac{n+\eta}{2})}{|V_{11}|^{\frac{p_2}{2}} |V_{22|1}|^{\frac{n+\eta}{2}} \Gamma_{p_2}(\frac{\eta}{2})}. \quad (\text{S.3})$$

Combining (S.1), (S.2), and (S.3) gives the desired result.

C. MCMC ALGORITHM

This section describes the MCMC sampling scheme for the full model described in Section 2.1 of the main paper. Since the component parameters are integrated out, the entire collection of parameters to be sampled in the MCMC is

$$\Theta = \left\{ \gamma, \phi, \lambda, \eta, \Psi, \alpha, \tilde{\mathbf{Z}} \right\}, \quad (\text{S.1})$$

where $\tilde{\mathbf{Z}}$ contains any latent element of \mathbf{Z} (i.e., are either missing data, or correspond to a discrete variable or censored observation).

The MCMC algorithm proceeds by performing Metropolis Hastings (MH) updates for each of the elements listed in Θ in a Gibbs fashion. The γ vector is updated with add/delete/swap proposals. The $\tilde{\mathbf{Z}}$ and Ψ are high dimensional and thus some creativity is needed to ensure good proposals. To accomplish this we first sample some component parameters from their conjugate distribution given the other parameters (for a fixed γ this is not difficult), and then use these component parameters to obtain good proposals for $\tilde{\mathbf{Z}}$ and Ψ , respectively. As mentioned in the main paper the ϕ_i can be updated individually with simple Gibbs sampling. However, this approach has known mixing issues and, thus, a modified split-merge algorithm (Jain and Neal, 2004) will be described below. The remaining updates for λ, η, α are more straight-forward random walk MH updates.

MH update for γ

The γ vector is updated with MH by proposing an add, delete, or swap move. That is, the proposal γ^* is generated as follows.

- (i) Set the proposal $\gamma^* = \gamma$
- (ii) Randomly choose an integer j^* from $1, \dots, p$.
- (iii) Flip the value of γ_{j^*} , i.e., $\gamma_{j^*}^* = 1 - \gamma_{j^*}$.
- (iv) If the set $\{j : \gamma_j \neq \gamma_{j^*}\}$ is not empty, draw a Bernoulli B^* with probability π .
- (v) If $B^* = 1$ randomly choose another j^{**} from the set $\{j : \gamma_j \neq \gamma_{j^*}\}$ and also set $\gamma_{j^{**}}^* = 1 - \gamma_{j^{**}}$, i.e., a swap proposal. If $B^* = 0$, leave γ^* as a single variable add/delete proposal.

Let $d(\gamma^* | \gamma)$ represent the density of this proposal. The MH ratio is then

$$MH = \frac{f(\mathbf{Z} | \gamma^*, \phi, \lambda, \eta, \Psi, \alpha) f(\gamma^*) d(\gamma | \gamma^*)}{f(\mathbf{Z} | \gamma, \phi, \lambda, \eta, \Psi, \alpha) f(\gamma) d(\gamma^* | \gamma)},$$

where $f(\mathbf{Z} | \gamma, \phi, \lambda, \eta, \Psi, \alpha)$ is the marginal likelihood provided in Result 3 and $f(\gamma)$ is the prior distribution for γ , i.e., independent Bernoulli(ρ). In the results of the main paper, ρ was set to 0.5.

MH update for α

The update for α was conducted via a MH random walk proposal on log scale. Draw a proposal $\log(\alpha^*) = \log(\alpha) + \epsilon$ for a deviate $\epsilon \sim N(0, s^2)$. The tuning parameter was set to $s = 1$ to achieve an acceptance rate $\approx 40\%$, and resulted in good mixing. Let the density of the proposal, given the current value of α be denoted $d(\alpha^* | \alpha)$. The only portion of the posterior that differs between the current value and the proposal is in the term

$$f(\phi | \alpha) = \prod_{i=2}^n \frac{n_{i,\phi_i} I_{\{n_{i,\phi_i} > 0\}} + \alpha I_{\{n_{i,\phi_i} = 0\}}}{i - 1 + \alpha} \propto \frac{\alpha^M \Gamma(\alpha)}{\Gamma(\alpha + n)}.$$

The MH ratio is then

$$MH = \frac{f(\phi | \alpha^*) f(\alpha^*) d(\alpha | \alpha^*)}{f(\phi | \alpha) f(\alpha) d(\alpha^* | \alpha)},$$

where $d(\alpha)$ is the density for a Gamma(A_α, B_α) random variable.

MH update for λ

The update for λ was conducted via a MH random walk proposal on log scale. Draw a proposal $\log(\lambda^*) = \log(\lambda) + \epsilon$ for a deviate $\epsilon \sim N(0, s^2)$. The tuning parameter was set to $s = 0.5$ to achieve an acceptance rate $\approx 40\%$. Let the density of the proposal, given the current value of λ be denoted $d(\lambda^* | \lambda)$. The MH ratio is then

$$MH = \frac{f(\mathbf{Z} | \gamma, \phi, \lambda^*, \eta, \Psi, \alpha) f(\lambda^*) d(\lambda | \lambda^*)}{f(\mathbf{Z} | \gamma, \phi, \lambda, \eta, \Psi, \alpha) f(\lambda) d(\lambda^* | \lambda)},$$

where $f(\lambda)$ is the density for a Gamma(A_λ, B_λ) random variable.

MH update for η

The update for η is entirely analogous to that for λ . A tuning parameter of $s = 1$ was used for η updates to encourage $\approx 40\%$ acceptance.

MH update for Ψ

The prior distribution is $\Psi \sim \mathcal{W}(\mathbf{P}, N)$. If the component parameters $\theta = \{\mu_{11}, \dots, \mu_{M1}, \Sigma_{111}, \dots, \Sigma_{M11}, \mathbf{b}_2, \mathbf{Q}_{21}, \mathbf{Q}_{22}\}$, were given then Ψ would have a conjugate update of the form,

$$\begin{aligned} \Psi_{22|1} | \theta &\sim \mathcal{W}(\mathbf{Q}_{22} + \mathbf{P}_{22}, N + \eta), \\ \Psi_{11} | \theta &\sim \mathcal{W}(\mathbf{P}^*, M\eta + N + p_2), \\ \Psi_{21} | \theta, \Psi_{11} &\sim \mathcal{MN}(-(\mathbf{P}_{22} + \mathbf{Q}_{22})(\mathbf{P}_{21} + \mathbf{Q}_{21})\Psi_{11}^{-1}, (\mathbf{P}_{22} + \mathbf{Q}_{22})^{-1}, \Psi_{11}), \end{aligned} \tag{S.2}$$

where

$$P^* = \left[P_{11}^{-1} + (P_{21} + Q_{21})'(P_{22} + Q_{22})^{-1}(P_{21} + Q_{21}) + P_{21}'P_{22}^{-1}P_{21} + Q_{21}'Q_{22}^{-1}Q_{21} + \sum_{m=1}^M \Sigma_{m11}^{-1} \right]^{-1},$$

and $\Psi_{22|1}$ is independent of Ψ_{11}, Ψ_{21} given θ .

We do not sample θ , so Ψ does not have such a conjugate update in the MCMC routine. However, we can generate a very good proposal Ψ^* in the following manner. Conditional on the current values of $\gamma, \phi, \lambda, \eta, \alpha, \tilde{Z}$, and Ψ , one could draw component parameters $\theta = \{\mu_{11}, \dots, \mu_{M1}, \Sigma_{111}, \dots, \Sigma_{M11}, b_2, Q_{21}, Q_{22}\}$, from their conjugate distribution provided in Section B. Conditional on the value of θ we could then draw from the distribution of $\Psi | \theta$ provided above. However, we do not want to have the current Ψ value involved in the update as this complicates the proposal density calculation. A simple fix is to draw component parameters θ^* conditional on the current values of $\gamma, \phi, \lambda, \eta, \alpha, \tilde{Z}$, but with Ψ fixed at some value $\tilde{\Psi}$, independent of the current (or previous) values in the chain. This way, the proposal density for Ψ^* is selected at random from a set of possible proposal distributions. The proposal density $d(\Psi^* | \Psi) = d(\Psi^*)$ is then conditional on θ^* and is simply the product of the densities in (S.2). This is allowable under the same principle used by Jain and Neal (2004) for the split-merge algorithm. Thus the MH ratio is then

$$MH = \frac{f(\tilde{Z} | \gamma, \phi, \lambda, \eta, \Psi^*, \alpha) f(\Psi^*) d(\Psi)}{f(\tilde{Z} | \gamma, \phi, \lambda, \eta, \Psi, \alpha) f(\Psi) d(\Psi^*)},$$

where $f(\Psi)$ is the density for a $\mathcal{W}(P, N)$ random variable. Note that this update would be made much easier if θ were just sampled in the MCMC as well. However, this makes it very difficult to update γ since the dimension of θ will be changing with γ . Reversible jump (RJ) MCMC could be used to overcome this issue by updating γ, θ jointly, but this comes with its own challenges. For a given γ and ϕ , however, drawing a θ to determine the proposal distribution as above poses no issues.

MH update for \tilde{Z}

The same logic used in the update of Ψ is used here as well. Conditional on the component parameters θ , the elements of \tilde{Z} have simple Gibbs updates. Namely, for a given observation i with missing values, the update for the missing z_i would be to draw from the normal distribution specified by ϕ_i and θ , conditional on the observed variables in z_i . If a value y_{ij} is discrete or censored, then the update for z_{ij} would be to draw from the normal distribution specified by ϕ_i and θ , conditional on the other variables in y_i and the conditional limits imposed by y_{ij} . Thus, a very similar trick as above is used. First divide \tilde{Z} up into K partitions and denote them $\tilde{Z}_1, \dots, \tilde{Z}_K$. We draw a separate θ_k^* for each partition by conditioning on the current values of $\gamma, \phi, \lambda, \eta, \alpha, \Psi$, and all \tilde{Z} values **except** \tilde{Z}_k . Update each of the elements of \tilde{Z}_k conditional on θ_k^* as described above to produce a proposal \tilde{Z}_k^* . Denote the density of this proposal as $d(\tilde{Z}_k^* | \tilde{Z}_k) = d(\tilde{Z}_k^*)$. The MH ratio is then,

$$MH = \frac{f(\tilde{Z}^* | \gamma, \phi, \lambda, \eta, \Psi, \alpha) d(\tilde{Z}_k^*)}{f(\tilde{Z} | \gamma, \phi, \lambda, \eta, \Psi, \alpha) d(\tilde{Z}_k)}.$$

MH update for ϕ

This is the most complex of the parameter updates as it uses a less standard split-merge MH approach (Jain and Neal, 2004) as this improves mixing dramatically over one-at-a-time Gibbs updates for the ϕ_i . However, the split-merge update does make use of the individual Gibbs update for the ϕ_i . This is provided

as

$$\begin{aligned} \text{pr}(\phi_i = m | \text{rest}) &\propto f(\mathbf{Z} | \gamma, \phi, \lambda, \eta, \Psi, \alpha) f(\phi_i | \phi_{-i}) \\ &\propto \begin{cases} \frac{n_{m(-i)} - 1}{n - 1 + \alpha} \left(\frac{n_{m(-i)} + \lambda}{n_m + \lambda} \right)^{\frac{p_1}{2}} \frac{|\Psi_{11}|^{\frac{n_{m(-i)} + \eta - p_2}{2}} \Gamma_{p_1}(\frac{n_m + \eta - p_2}{2})}{|\mathbf{V}_{m11}|^{\frac{n_m + \eta - p_2}{2}} \Gamma_{p_1}(\frac{n_{m(-i)} + \eta - p_2}{2})} & \text{if } m = \phi_l \text{ for some } \phi_l \in \phi_{-i}, \\ \frac{\alpha}{n - 1 + \alpha} \left(\frac{\lambda}{\lambda + 1} \right)^{\frac{p_1}{2}} \frac{|\Psi_{11}|^{\frac{\eta - p_2}{2}} \Gamma_{p_1}(\frac{\eta + 1 - p_2}{2})}{|\mathbf{V}_{m11}|^{\frac{1 + \eta - p_2}{2}} \Gamma_{p_1}(\frac{\eta - p_2}{2})} & \text{for } m = M + 1, \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \quad (\text{S.3})$$

where ϕ_{-i} is the ϕ vector with out the i^{th} element and ϕ_{-i} has been relabeled if necessary so that it has at least one $\phi_l = m$ for $m = 1, \dots, M$. The split-merge MH update then works as follows.

1. Set $\phi^* = \phi$. Select two points i and i' at random. Let $\mathcal{C} = \{l : \phi_l = \phi_i \text{ or } \phi_l = \phi_{i'}\}$.
- 2.(a) If $\phi_i = \phi_{i'}$, then propose a split move to divide \mathcal{C} into two groups in ϕ^* .
 - (i) For $l \in \mathcal{C}$, set $\phi_l^{\text{launch}} = \begin{cases} \phi_i & \text{if } \|\mathbf{z}_l^{(1)} - \mathbf{z}_i^{(1)}\| \leq \|\mathbf{z}_l^{(1)} - \mathbf{z}_{i'}^{(1)}\|, \\ M + 1 & \text{otherwise} \end{cases}$
 - (ii) Conduct a Gibbs update sweep (S.3) to all $\phi_l : l \in \mathcal{C}$, restricted to $\phi_l = \phi_i$ or $\phi_l = M + 1$.
 - (iii) Repeat step (iii) for a total of L passes through \mathcal{C} . This determines ϕ^{launch} and the randomly chosen proposal distribution to be used next in step 2(a)(iv).
 - (iv) Set $\phi^* = \phi^{\text{launch}}$ and conduct one further restricted Gibbs sweep to the $\phi_l^{\text{launch}} : l \in \mathcal{C}$. The proposal density $d(\phi^* | \phi)$ is the product of the restricted Gibbs sampling probabilities in this final sweep, whereas $d(\phi | \phi^*) = 1$.
- (b) If $\phi_i \neq \phi_{i'}$, then propose a merge move to combine the observations in \mathcal{C} into one group ϕ^* .
 - (i) Set $\phi_l^* = \phi_i$ for all $l \in \mathcal{C}$. The proposal density is $d(\phi^* | \phi) = 1$.
 - (ii) Conduct steps 2(a)(i) - 2(a)(iv) in order to evaluate the reverse proposal density $d(\phi | \phi^*)$.
 - (iii) The reverse proposal density $d(\phi | \phi^*)$ is the product of restricted Gibbs sampling probabilities for moving from ϕ^{launch} to ϕ .
3. The MH ratio is then,

$$\begin{aligned} MH &= \frac{f(\mathbf{Z} | \gamma, \phi^*, \lambda, \eta, \Psi, \alpha) f(\phi^*) d(\phi | \phi^*)}{f(\mathbf{Z} | \gamma, \phi, \lambda, \eta, \Psi, \alpha) f(\phi) d(\phi^* | \phi)} \\ &= \frac{\prod_{m \in \{\phi_l^* : l \in \mathcal{C}\}} \left[\alpha(n_m^* - 1)! \left(\frac{\lambda}{n_m^* + \lambda} \right)^{\frac{p_1}{2}} \frac{|\Psi_{11}|^{\frac{\eta - p_2}{2}} \Gamma_{p_1}(\frac{n_m^* + \eta - p_2}{2})}{|\mathbf{V}_{m11}^*|^{\frac{n_m^* + \eta - p_2}{2}} \Gamma_{p_1}(\frac{\eta - p_2}{2})} \right] d(\phi | \phi^*)}{\prod_{m \in \{\phi_l : l \in \mathcal{C}\}} \left[\alpha(n_m - 1)! \left(\frac{\lambda}{n_m + \lambda} \right)^{\frac{p_1}{2}} \frac{|\Psi_{11}|^{\frac{\eta - p_2}{2}} \Gamma_{p_1}(\frac{n_m + \eta - p_2}{2})}{|\mathbf{V}_{m11}|^{\frac{n_m + \eta - p_2}{2}} \Gamma_{p_1}(\frac{\eta - p_2}{2})} \right] d(\phi^* | \phi)}, \end{aligned}$$

where n_m is the number of $\phi_l = m$ and n_m^* is the number of $\phi_l^* = m$. Draw a Uniform(0,1) and accept or reject in the usual manner on the basis of the MH ratio.

4. Perform one final (unrestricted) Gibbs update over *all* observations, i.e., for each $\phi_l, l = 1, \dots, n$. As discussed in Jain and Neal (2004), alternating between split-merge and Gibbs updates produces an ergodic Markov chain.

Joint MH update for γ and ϕ

The MCMC routine then consists of applying each of the above updates in turn to complete a single MCMC iteration, with the exception that the γ update be applied L_g times each iteration. Also, as discussed in the main paper, to improve mixing we recommend using the following joint update for γ and ϕ in place of the individual updates for γ and ϕ every other iteration (or simply in addition to them every iteration). This is fairly straight-forward as the proposals are generated by simply generating a γ^* as

above, then a ϕ^* conditional on γ^* as above. The MH ratio for such an update is then,

$$MH = \frac{f(\mathbf{Z} | \gamma^*, \phi^*, \lambda, \eta, \Psi, \alpha) f(\gamma^*) f(\phi^*) d(\gamma | \gamma^*) d(\phi | \phi^*, \gamma)}{f(\mathbf{Z} | \gamma, \phi, \lambda, \eta, \Psi, \alpha) f(\gamma) f(\phi) d(\gamma^* | \gamma) d(\phi^* | \phi, \gamma^*)},$$

By the same rational as that used in Jain and Neal (2004), both the individual updates and the joint update leave the possibility for the states to remain unchanged, therefore applying each of these transitions in turn will produce an ergodic Markov chain.

D. MCMC TRACE PLOTS

In order to get a big picture view of the mixing of the MCMC algorithm, the MCMC trace plots for the number of informative variables p_1 and the number of clusters M are provided below in Figure S.1 for two separate MCMC chains (in blue and red respectively) of 75,000 iterations each. It is apparent the the mixing is slower for the variable selection than for the cluster membership, however, over 75,000 iterations, if a chain spends a non-negligible number of iterations at a value for p_1 (i.e., 4, 5, or 6), then their are dozens of switches to that model size that occur all throughout the life of the chains. A more granular view of the mixing is also provided by the MCMC trace plots for the individual γ_j (for $j = 1, \dots, 46$) in Figure S.2 and each of the ϕ_i (for $i = 1, \dots, 96$; the exhaustive list of all 487 subjects trace plots all looked similar to these) in Figure S.3. Once again results are provided for two chains. Mixing is slow for γ_j , thus the need for so many MCMC iterations. Over 75,000 iterations, if a chain spends a non-negligible number of iterations at either 0 or 1 for a given variable, then their are generally many switches that occur all throughout the life of the chains. Mixing for the ϕ_i on the other hand is quite good in comparison; all observations that spend a non-negligible number of iterations with more than one cluster switch back and forth between cluster memberships quite regularly. While there were a maximum of 12 clusters observed over all MCMC iterations from both chains, labels 7 - 12 only accounted for a total of 0.0036 of the posterior probability, so only labels 1-6 are displayed for clearer presentation. The ϕ_i displayed in these plots are **not** the raw ϕ_i that are subject to label switching. Rather, they have been relabeled for clearer interpretation according to the information theoretic approach discussed in Section 2-6 so that, for instance, a label of 1 can be interpreted as “belonging to the same cluster 1” regardless of the MCMC iteration number.

Fig. S.1: MCMC Trace plots for p_1 and m

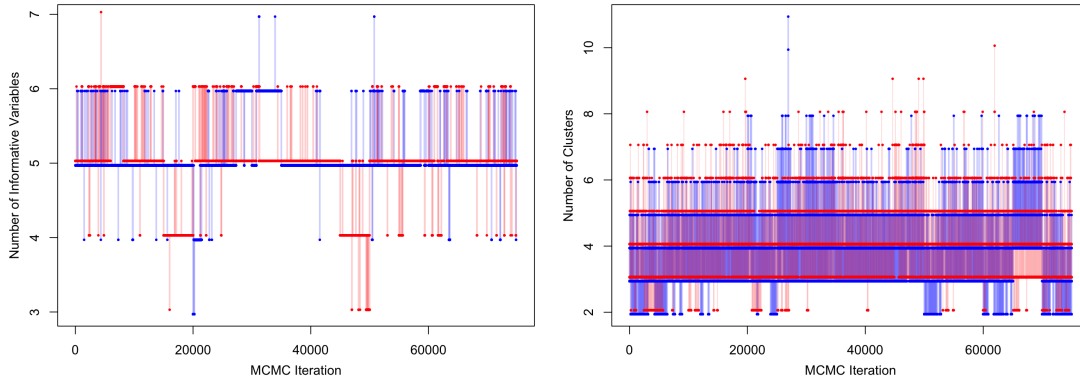


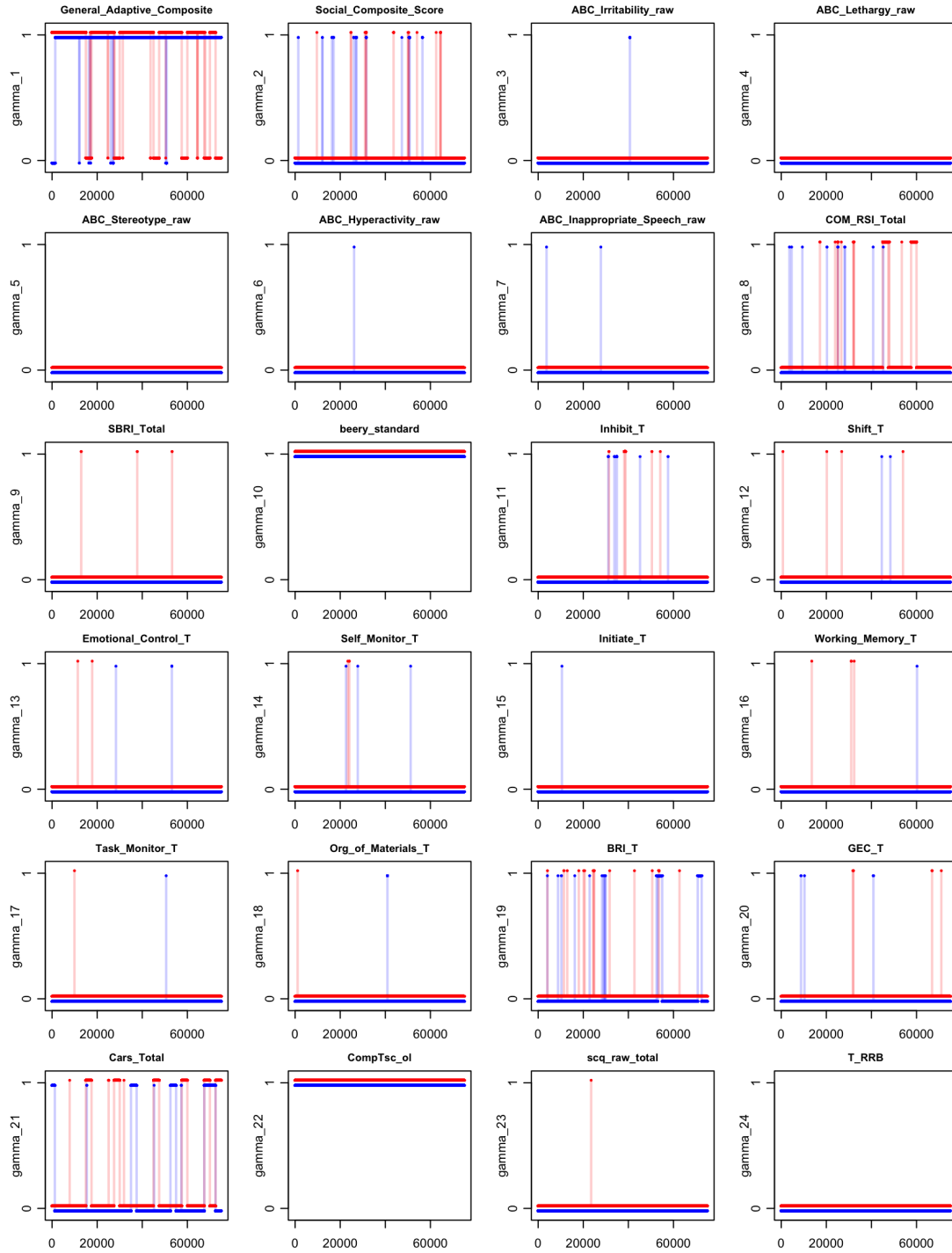
Fig. S.2: MCMC Trace plots for γ_j 

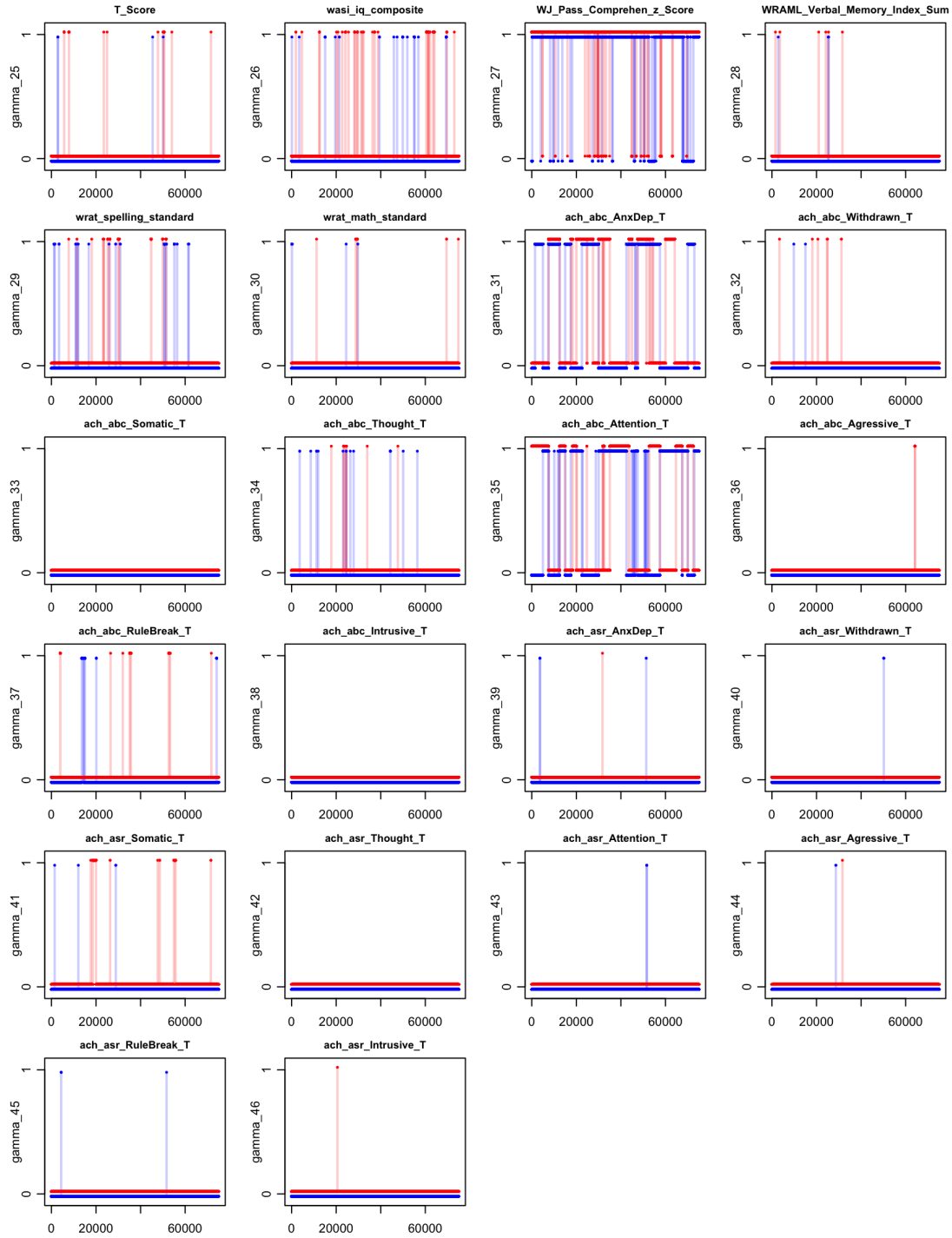
Fig. S.2: MCMC Trace plots for γ_j 

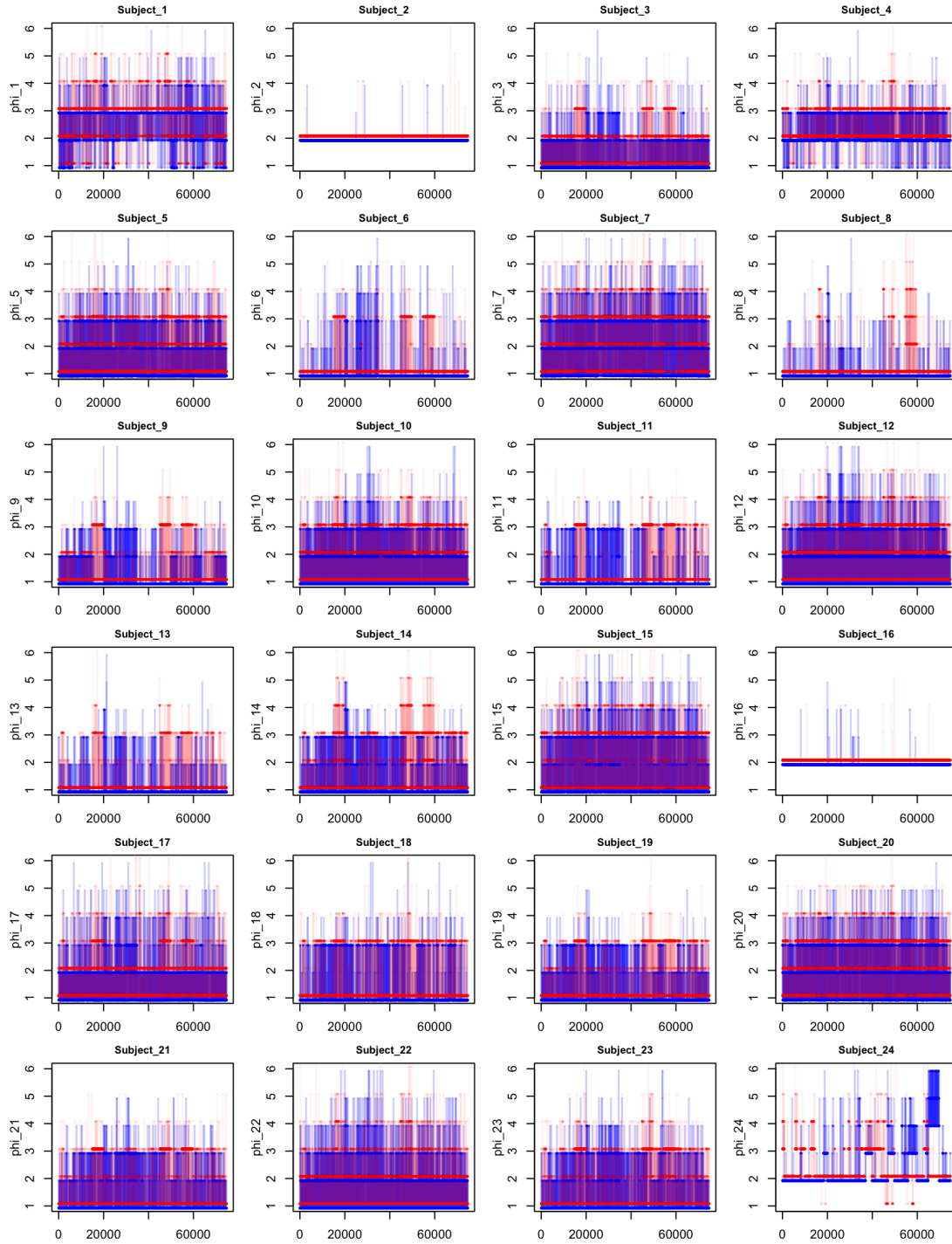
Fig. S.3: MCMC Trace plots for ϕ_i 

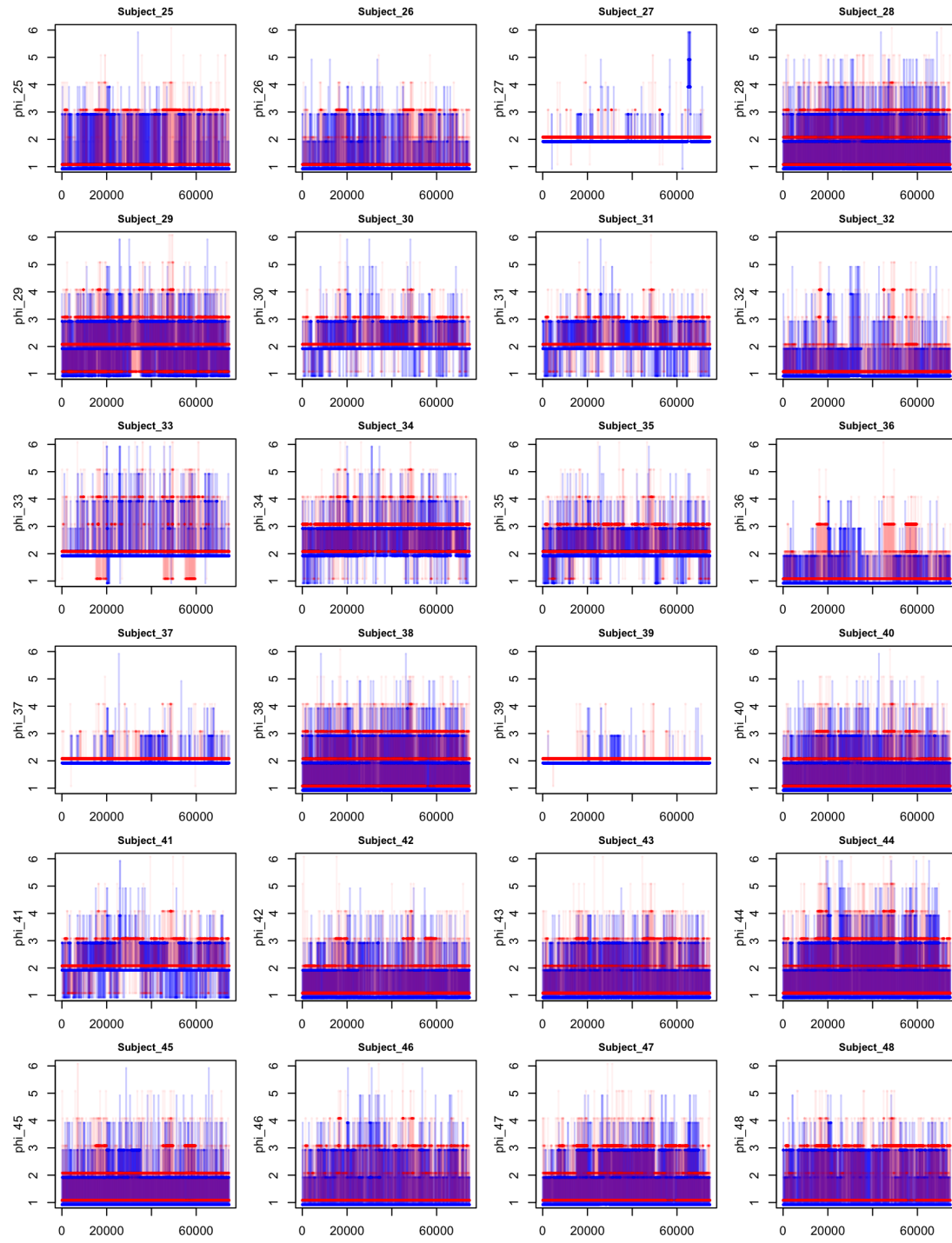
Fig. S.2: MCMC Trace plots for ϕ_i 

Fig. S.2: MCMC Trace plots for ϕ_i

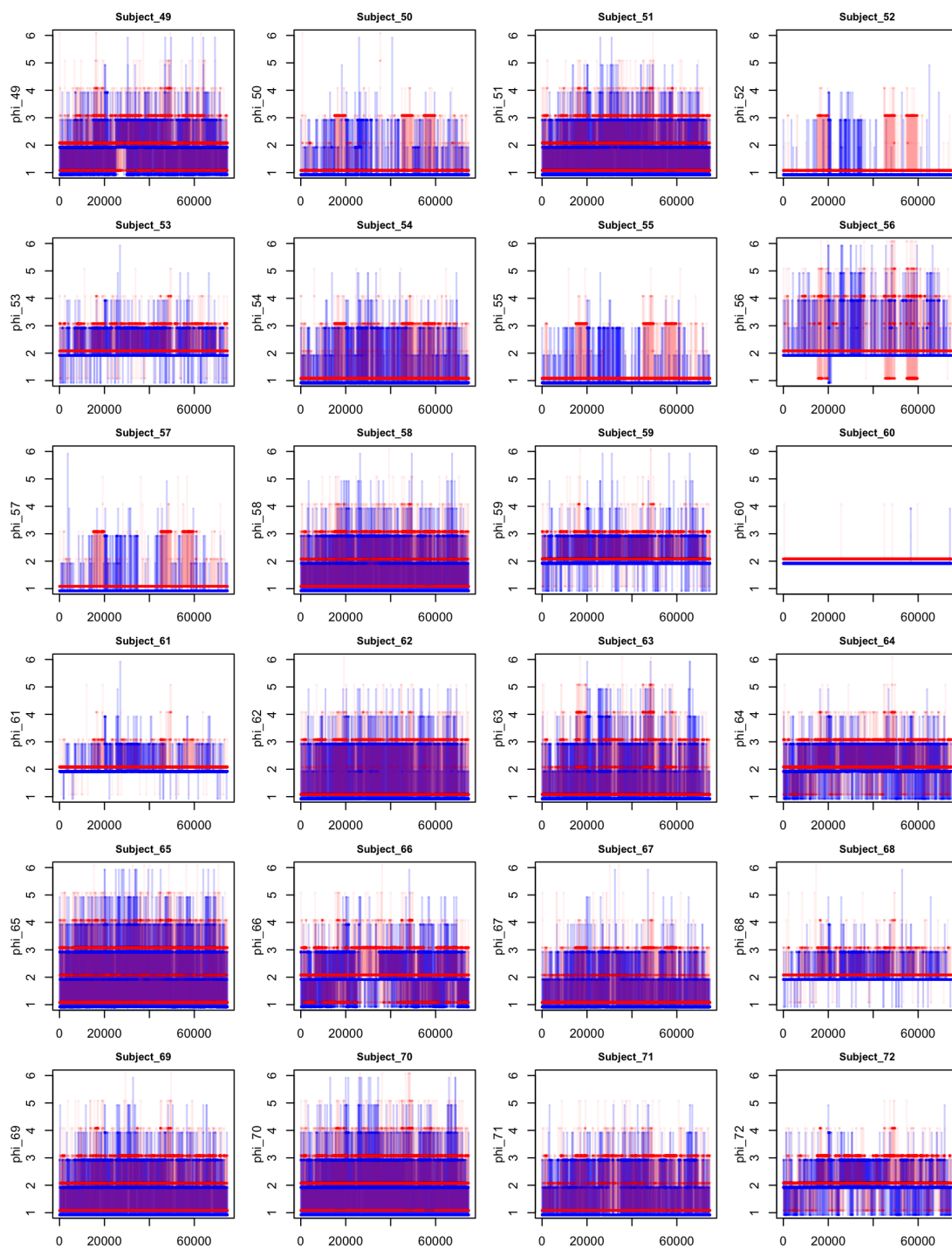


Fig. S.2: MCMC Trace plots for ϕ_i 